

AD658894

AMERICAN MATHEMATICAL SOCIETY

Lecture Notes Prepared in Connection with the Summer Seminar

on

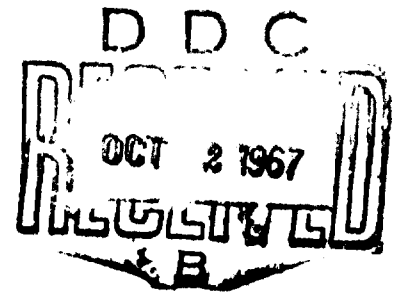
Mathematics of the Decision Sciences

held at

Stanford University

Stanford, California

July 10 - August 11, 1967



Sponsored by: Atomic Energy Commission, Contract # AT(30-1)-3164 Modification No. 3  
Air Force Office of Scientific Research (with NIH), Contract #PH-43-67-712  
Army Research Office (Durham), Contract #DA-31-124-ARO(D)-82  
Office of Naval Research, Grant # Nonr(G)-00003-67  
National Institutes of Health (with AFOSR), Contract #PH-43-67-712  
National Science Foundation, Grant #GZ-403

Informally distributed manuscripts and articles should be treated as a personal communication and are not for library use. Reference to the contents in any publication should have the author's prior approval.

This document has been approved  
for public release and under the  
distribution is unlimited.

PAGES \_\_\_\_\_  
ARE  
MISSING  
IN  
ORIGINAL  
DOCUMENT

## TABLE OF CONTENTS

- R. M. Thrall  
Survey of Mathematical Programming
- E. Polak  
Necessary Conditions of Optimality in Control and Programming
- George B. Dantzig  
Mathematical Programming
- Michel L. Balinski  
Survey of Mathematical Programming
- Terry Rockafellar  
Nonlinear Programming
- Kenneth Arrow  
Mathematical Economics
- David Gale  
Mathematical Economics
- J. B. Rosen  
Computational Aspects of Control Theory
- A. W. Tucker  
Mathematical Programming
- Richard Cottle  
Mathematical Programming
- D. R. Fulkerson  
Networks and Graphs
- Jack Edmonds  
Combinatorial Methods
- Ralph E. Gomory  
Integer Programming
- Carlton Lemke  
Mathematical Programming
- Arthur F. Veinott, Jr.  
Optimal Inventory Control
- Donald L. Iglehart  
Diffusion Approximations in Applied Probability
- Herman Chernoff  
Optimal Stochastic Control

1001  
Richard E. Barlow  
Reliability Theory

Cyrus Derman  
Markovian Decision Processes

M. Frank Norman  
Learning Theory

David Krantz  
Measurement and Psychophysics

Abraham Taub  
Computer Science

Andrzej J. Ehrenfeucht  
Perception Problems

LECTURES DELIVERED DURING THE SEMINAR BUT NOT  
APPEARING IN THIS VOLUME

Alan Hoffman  
Mathematical Programming

Samuel Karlin  
Branching Processes

Victor Klee  
Convexity

Harold Kuhn  
Mathematical Economics

William Miller  
Computer Science

Lucien Neustadt  
Control Theory

Stanley Peters  
Mathematical Linguistics

Herbert Robbins  
Mathematical Statistics

Philip Wolfe  
Nonlinear Programming

D. O. Siegmund

**SURVEY OF MATHEMATICAL PROGRAMMING**

by

**R. M. THRALL**

at the

**American Mathematical Society Summer Seminar**

on the

**Mathematics of the Decision Sciences**

**Stanford University**

**July - August 1967**

## 1. Introduction

Optimization problems have been part of mathematics from its earliest days. The general constrained optimization problem can be written in form

$$\min \phi(X) \text{ for } X \in P$$

where  $P$  is some set and  $\phi$  is a (real valued) scalar function whose domain contains  $P$ . Linear programming refers to the special case in which  $P$  is a polyhedral subset of a vector space and  $\phi$  is linear.

Let  $A = || a_{ij} ||$  be a  $p$ -by- $m$  matrix, let  $X = || x_i ||$ ,  $C = || c_i ||$  be  $m$ -by-1 vectors, let  $B$  be a  $p$ -by-1 vector and let  $d$  be a scalar. [In these lectures all scalars will be real numbers.] Then the linear program

$$(1.1) \quad \min Z = d + C^T X$$

subject to the constraints

$$\begin{aligned} AX &= B \\ X &\geq 0 \end{aligned}$$

is said to be in standard form. We sometimes call  $A$  the coefficient matrix,  $X$  the activity vector,  $C$  the cost vector,  $B$  the constraint vector, and  $d$  the fixed cost or initial cost. Then

$$P = \{ X \mid AX = B, X \geq 0 \}$$

is called the set of feasible vectors. If  $P$  is empty, the problem is said to be infeasible.

A problem in standard form can be presented as the tableau matrix

(1.2)

$M =$

-d	$C^T$
B	A

or in more detail

( 2 )

(1.3)

$$z \quad 1 = x_1 \quad x_j \quad x_k \quad x_n$$

$z$	$-d$	$c_1 \dots c_j \dots c_k \dots c_n$
$b_1$	$a_{11} \dots a_{1j} \dots a_{1k} \dots a_{1n}$	
$\vdots$	$\vdots$	$\vdots$
$b_i$	$a_{i1} \dots a_{ij} \dots a_{ik} \dots a_{in}$	
$\vdots$	$\vdots$	$\vdots$
$b_h$	$a_{h1} \dots a_{hj} \dots a_{hk} \dots a_{hn}$	
$\vdots$	$\vdots$	$\vdots$
$b_p$	$a_{p1} \dots a_{pj} \dots a_{pk} \dots a_{pn}$	

The zeroth (top) row is read as

$$z - d = c_1 x_1 + \dots + c_n x_n$$

and the i-th row as

$$b_i = a_{i1}x_1 + \dots + a_{in}x_n \quad (i = 1, \dots, p).$$

As a notational convenience we sometimes write

(1.4)  $a_{00} = -d, a_{0j} = c_j \quad (j = 1, \dots, n), a_{i0} = b_i \quad (i = 1, \dots, p)$

For example,

(1.5)

0	-8	-19	-7	0	0
25	3	4	1	1	0
50	1	3	3	0	1

describes a program in standard form, having  $p = 2$  and  $m = 5$ .

The basic computational step involved in the simplex algorithm for solving linear programming problems is the pivot operation of Gaussian elimination. To pivot on a position  $(h,k)$  we must have the pivot element  $a_{hk}$  different from zero. The pivot operation has two steps:

Normalize: Divide the pivot row (row  $h$ ) by the pivot element;  
and  
Sweep out: Subtract suitable multiples of the pivot row from the remaining rows so as to obtain zeros in all positions of the pivot column (column  $k$ ) except for the 1 in the pivot position  $(h,k)$ .

Thus after pivoting on position  $(h,k)$  the matrix  $M$  of (1.2) (or (1.3)) becomes

$$(1.6) \quad M^* = \begin{array}{c} \begin{array}{cc} & \begin{array}{cc} j & k \end{array} \\ \begin{array}{c} i \\ h \end{array} & \begin{array}{cc} a_{ij} - a_{ik}a_{hj}/a_{hk} & 0 \\ a_{hj}/a_{hk} & 1 \end{array} \end{array} \end{array}$$

or

$$(1.7) \quad \begin{aligned} a_{hj}^* &= a_{hj}/a_{hk} \quad (j = 0, 1, \dots, n) \\ a_{ij}^* &= a_{ij} - a_{ik}a_{hj}/a_{hk} \quad (i = 0, 1, \dots, p; i \neq h; j = 0, 1, \dots, n) \end{aligned}$$

For example, if we pivot on position  $(1,2)$  in (1.5) we get

$$(1.8) \quad \begin{array}{c|ccccc} 475/4 & 25/4 & 0 & -9/4 & 19/4 & 0 \\ \hline 25/4 & 3/4 & 1 & 1/4 & 1/4 & 0 \\ \hline 125/4 & -5/4 & 0 & 9/4 & -3/4 & 1 \end{array}$$

Clearly, pivoting on position  $(h,k)$  represents solving equation  $h$  for  $x_k$  and using this equation to eliminate  $x_h$  from the remaining equations.

Lemma 1A. Let

$$(1.9) \quad M^* = \begin{array}{|c|c|} \hline -d^* & C^{*T} \\ \hline B^* & A^* \\ \hline \end{array}$$

be obtained from the  $M$  of (1.2) by a pivot operation on position  $(h,k)$ . Then we have

$$(1.10) \quad M^* = QM$$

where

$$(1.11) \quad Q = \begin{bmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & & -a_{hk}/a_{hk} & & \\ & & & \vdots & & \\ & & & 1/a_{hk} & 1 & \\ & & & \vdots & & \\ & & & -a_{pk}/a_{hk} & & 1 \end{bmatrix}$$

( 4 )

is a non-singular  $(p + 1) - \text{by} - (p + 1)$  matrix differing from identity only in column  $h$ , i.e.

$$q_{ij} = \delta_{ij} \text{ (Kronecher's delta function)}$$

$$i = 0, 1, \dots, p; j \neq k$$

(1.12)

$$q_{0h} = -c_h/a_{hk}$$

$$q_{ih} = -a_{ik}/a_{hk} \quad i = 1, \dots, p; i \neq k$$

$$q_{hh} = 1/a_{hk}.$$

Moreover, the problems described by  $M$  and  $M^*$  have the same feasible sets  $P$  and  $P^*$  and for any feasible vector  $X$  we have

$$d^* + c^{*T}X = d + c^T X.$$

Theorem 1B. Let  $M^*$  be obtained from  $M$  by a sequence of pivot operations on positions not in the zeroth row or column. Then (1.10) holds for a nonsingular matrix  $Q$  having the form

$$Q = \begin{bmatrix} 1 & R \\ 0 & P \end{bmatrix},$$

i.e., the zero-th column of  $Q$  is the initial unit vector. The linear programs defined by  $M^*$  and  $M$  are equivalent in the sense that every constraint equation of either problem is a linear combination of constraint equations of the other and that for any vector  $X$  which satisfies the constraint equations  $AX = B$  or, equivalently,  $A^*X = B^*$  we have also  $d + C^T X = d^* + C^{*T} X$ .

Note: the last statement is formulated to include vectors  $X$  which may be infeasible in the sense of having some negative components.

Proof: Using the theory of partitioned matrices we observe that a product of matrices of the form (1.11) each having  $h \geq 1$  and  $k \geq 1$  has the form (1.13). Moreover,  $Q^{-1}$  also has the form (1.13). The conclusions concerning equivalence then follow from (1.10). In particular we observe that

$$||-d^*c^{*T}|| = ||-dC^T|| + R||BA||,$$

i.e., the cost rows differ by a linear combination of constraint rows. Now

( 5 )

$$d^* + C^{*T}X = \begin{bmatrix} -d^*C^{*T} \\ \bar{X}^1 \end{bmatrix} = \begin{bmatrix} -dC^T \\ + R \end{bmatrix} \begin{bmatrix} BA \\ \bar{X}^1 \end{bmatrix} =$$

$$d + C^T X + R \bar{X}^1 = d + C^T X.$$

For later use we note that if

$$(1.14) \quad Q^{-1} = \begin{bmatrix} 1 & R^* \\ 0 & P^* \end{bmatrix},$$

then

$$(1.15) \quad P^* = P^{-1}, R^* = -RP^{-1}.$$

The conclusion of theorem 1B states that a sequence of pivots leads from any linear programming problem M to a new one M\* which is equivalent to it in a very strong sense of equivalence.

## 2. Canonical form, Convergence

We introduce a notation for submatrices which will be useful in what follows.

Let  $A$  be a  $p$ -by- $n$  matrix, let  $C$  be an  $n$ -by-1 column vector, and let  $R = (r_1, \dots, r_q)$  be any sequence of integers with  $1 \leq r_i \leq n$ . Then we denote by  $A_R$  the  $p$ -by- $q$  matrix whose columns in order are  $A_{r_1}, \dots, A_{r_q}$ , and we denote by  $C_R$  the  $q$ -by-1 column vector whose components in order are  $c_{r_1}, \dots, c_{r_q}$ , i.e.,

$$(2.1) \quad A_R = \begin{vmatrix} | & | & A_{r_1} & \dots & A_{r_q} & | \\ | & | & & & & | \end{vmatrix}, \quad C_R = \begin{vmatrix} | & | & c_{r_1} & | \\ | & | & \cdot & | \\ | & | & \cdot & | \\ | & | & \cdot & | \\ | & | & c_{r_q} & | \end{vmatrix}.$$

For example, let

$$A = \begin{vmatrix} | & | & 1 & 0 & -2 & 1 & 1 & 0 & 2 & | \\ | & | & 2 & 1 & 1 & 0 & -2 & 0 & 2 & | \\ | & | & -3 & 0 & 2 & 0 & 1 & 1 & 3 & | \end{vmatrix}, \quad B = \begin{vmatrix} | & | & 3 & | \\ | & | & 7 & | \\ | & | & 2 & | \end{vmatrix}, \quad C = \begin{vmatrix} | & | & 3 & | \\ | & | & 0 & | \\ | & | & -2 & | \\ | & | & 0 & | \\ | & | & -4 & | \\ | & | & 0 & | \\ | & | & 7 & | \end{vmatrix},$$

$$d = 14,$$

and let  $R = (2, 7, 3, 1)$ ; then

$$A_R = \begin{vmatrix} | & | & 0 & 2 & -2 & 1 & | \\ | & | & 1 & 2 & 1 & 2 & | \\ | & | & 0 & 3 & 2 & -3 & | \end{vmatrix}, \quad C_R = \begin{vmatrix} | & | & 0 & | \\ | & | & 7 & | \\ | & | & -2 & | \\ | & | & 3 & | \end{vmatrix}$$

and for  $S = (4, 2, 6)$ ,

$$A_S = \begin{vmatrix} | & | & 1 & 0 & 0 & | \\ | & | & 0 & 1 & 0 & | \\ | & | & 0 & 0 & 1 & | \end{vmatrix}, \quad C_S = \begin{vmatrix} | & | & 0 & | \\ | & | & 0 & | \\ | & | & 0 & | \end{vmatrix}.$$

Let

$$(2.2) \quad S = (s_1, \dots, s_p)$$

be an ordered sequence of  $p$  different integers selected from the set  $\{1, 2, \dots, n\}$ . The problem (1.1) is said to be in canonical form with respect to the basic sequence  $S$  if

- (C1)  $A_S = I_p$ , canonical coefficients,
- (C2)  $C_S = 0$ , canonical costs,
- (C3)  $B \geq 0$ , canonical constants.

Note that C1 states that the system of equations has been solved for  $x_{s_1}, \dots, x_{s_p}$  (the basic variables) in terms of the remaining (non-basic) variables; C2 states that the equations have been used to eliminate the basic variables from the cost function; and C3 states that in the solved form of the equations the constant terms are non-negative.

Associated with a problem in canonical form, there is a feasible vector  $x^S$  and corresponding cost  $z^S$  given by

$$(2.3) \quad \begin{aligned} x_{s_i}^S &= b_i \quad (i = 1, \dots, p) \\ x_g^S &= 0 \quad g \notin S \\ z^S &= d. \end{aligned}$$

This solution is referred to as the basic solution associated with the basic sequence  $S$ .

If we let  $G$  denote the sequence remaining when the elements of  $S$  are deleted from  $(1, \dots, n)$  then (2.3) can be written as

$$(2.3') \quad x_S^S = B, \quad x_G^S = 0, \quad z^S = d.$$

The example (1.5) is canonical with respect to the basic sequence  $S = (4, 5)$ . Here  $G = (1, 2, 3)$  and the associated basic solution and cost are

$$x_S^S = \begin{bmatrix} x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 25 \\ 50 \end{bmatrix}, \quad x_G^S = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad z^S = 0.$$

A vector  $F = ||f_1 \ f_2 \ \dots \ f_n||$  is said to be lexicographically positive, written  $F \succ 0$ , if the first non-zero component of  $F$  is positive, i.e., if for some  $i$ ,  $1 \leq i \leq n$ , we have  $f_1 = \dots = f_{i-1} = 0$ ,  $f_i > 0$ . For example,  $||2 \ -1 \ 3||$ ,  $||0 \ 0 \ 1||$ , and  $||0 \ 2 \ -3 \ 4||$  are all lexicographically positive whereas  $||0 \ -1 \ 7||$ ,  $||0 \ 0 \ 0||$ , and  $||-1 \ 7 \ 8||$  are not. If  $F - G \succ 0$  we say that  $F$  is lexicographically greater than  $G$ , written  $F \succ G$ .

If, for a problem in canonical form, the strict inequality  $B \succ 0$  holds, then we have automatically that:  $C^4$ --every row of  $||B \ A||$  is lexicographically positive. If some components of  $B$  are zero but the first non-zero  $a_{ij}$  is each such row is positive then  $C^4$  holds. If, in particular,  $S = (1, 2, \dots, p)$ , then  $C^4$  is a consequence of  $C^3$  and  $C^1$ . More generally, if  $C^1$  and  $C^3$  hold, a permutation of columns  $1, \dots, n$  will yield  $C^4$ .

Theorem 2A. Consider a linear programming problem in canonical form. Then one of the following alternatives holds:

- (i)  $C \geq 0$  and the associated basic solution is optimal.
- (ii) There exists  $k$  such that  $c_k < 0$  and  $A_k \leq 0$  and the cost function has no lower bound.
- (iii) There exists  $k$  such that  $c_k < 0$  and  $i$  such that  $a_{ik} > 0$ , and if  $C^4$  holds there is a position in column  $k$  at which a pivot transformation results in an equivalent problem in canonical form, satisfying  $C^4$ , and for which the zero-th row  $||-d^* \ c^*||$  is lexicographically greater than that  $||-d \ C||$  of the initial problem.

Proof: To establish (i) we note that for any feasible  $X$  we have

$$(2.4) \quad Z = d + C^T X = d + C_S^T X_S + C_G^T X_G = d + C_G^T X_G \geq d = Z^S.$$

The inequality follows from the fact that  $C_G \geq 0$  and  $X_G \geq 0$ ; hence  $X^S$  is optimal and  $Z^S$  is the minimal cost.

The matrix  $M$  in (2.5) illustrates case (i).

$$(2.5) \quad M = \begin{array}{c|ccccc} & -20 & 0 & 0 & 3 & 4 & 5 \\ \hline 9 & 1 & 0 & 3 & -2 & 7 \\ 2 & 0 & 1 & 3 & -3 & -2 \end{array}$$

To establish (ii) consider the equations

$$(2.6) \quad \begin{aligned} b_i &= x_{s_i} + a_{i_k} x_k \quad i = 1, \dots, p \\ z &= d + c_k x_k \end{aligned}$$

obtained from (1.1) by setting all of the non-basic variables equal to zero except for  $x_k$ . As  $x_k$  takes on larger and larger positive values, each  $x_{s_i}$  either grows larger also (if  $a_{i_k} < 0$ ) or stays constant (if  $a_{i_k} = 0$ ) and  $z$  either is initially or ultimately becomes negative and takes on larger and larger negative values. Thus  $z \rightarrow -\infty$  as  $x_k \rightarrow \infty$  and always with feasible vectors  $X$ . The situation is pictured in Figure 2.1.

The matrix  $M$  in (2.7) illustrates case (ii) with  $k = 4$ .

$$(2.7) \quad M = \begin{array}{c|ccccc} & -20 & 0 & 0 & 3 & -4 & 5 \\ \hline & 9 & 1 & 0 & 3 & -2 & 7 \\ & 2 & 0 & 1 & 3 & -3 & -2 \end{array}$$

Case (iii) differs from case (ii) since for  $a_{i_k} > 0$ , equations (2.6) place an upper bound of  $b_i/a_{i_k}$  on  $x_k$ . See Figure 2.2.

For feasibility, we must have

$$(2.8) \quad x_k \leq \min \{ b_i/a_{i_k} \mid a_{i_k} > 0 \} = b_h/a_{h_k}.$$

If now we pivot at  $(h,k)$ , the new basic solution is given by (2.6) with  $x_k = b_h/a_{h_k}$ , and if  $b_h > 0$ , the cost  $Z$  is reduced by addition of the negative amount  $b_h c_k/a_{h_k}$ ; i.e.,

$$(2.9) \quad z' = z - b_h c_k/a_{h_k}.$$

The matrix in (2.10) illustrates case (iii) with  $k = 3$ ,  $h = 2$ :

$$(2.10) \quad M = \begin{array}{c|ccccc} & -20 & 0 & 0 & -3 & 4 & 5 \\ \hline & 9 & 1 & 0 & 3 & -2 & 7 \\ & 2 & 0 & 1 & 3 & 3 & -2 \end{array} \quad \begin{array}{l} b_i/a_{i_k} \\ 9/3 \\ 2/3 \end{array} ;$$

and after pivoting on position (2,3) we obtain

$$(2.11) \quad M = \begin{array}{c|ccccc} & -18 & 0 & 1 & 0 & 7 & 3 \\ \hline & 7 & 1 & -1 & 0 & -5 & 9 \\ & 2/3 & 0 & 1/3 & 1 & 1 & -2/3 \end{array}$$

( 10 )

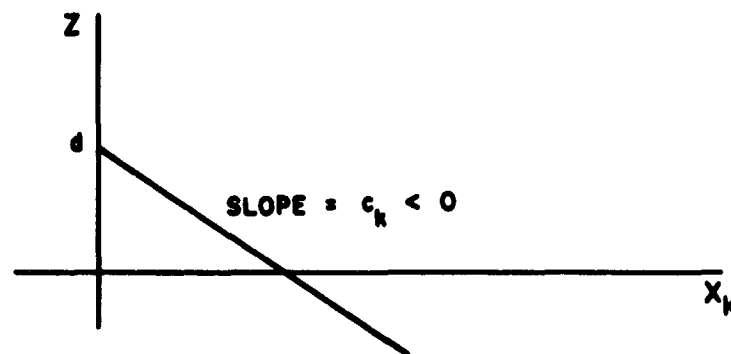
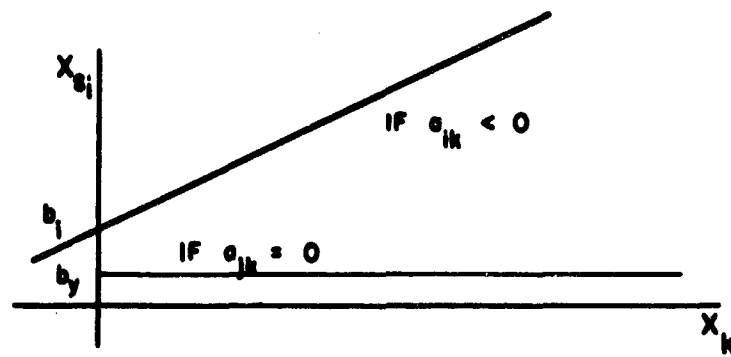


Figure 2.1. Geometry of Unbounded Case.

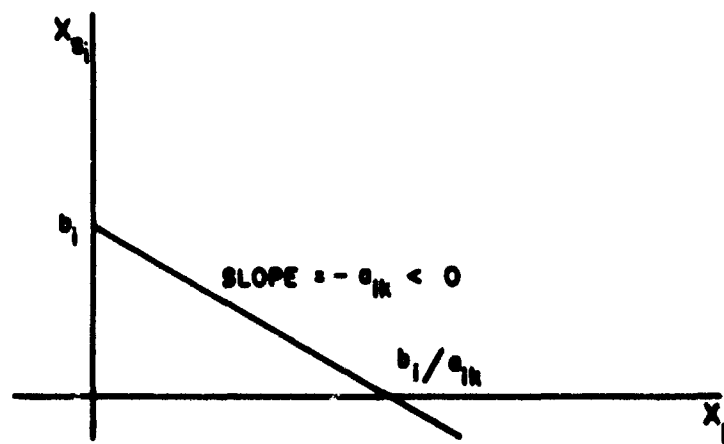


Figure 2.2. Geometry of Bounded Case.

which is in case (i) and therefore the associated basic solution

$$(2.12) \quad X = \begin{bmatrix} 7 \\ 0 \\ 2/3 \\ 0 \\ 0 \end{bmatrix}, \quad z = 18$$

is optimal.

In this example we had  $B > 0$ , and hence reduction in the cost  $z$  (from 20 to 18). It sometimes happens that  $b_h = 0$  and then (cf. (2.9)) there is no improvement in  $z$ ; this is the phenomena called degeneracy. To handle degeneracy we look not just at  $-d^*$  but at the entire row  $||-d^* c^{*T}||$ . For it we have

$$(2.13) \quad ||-d^* c^{*T}|| = ||-d c^T|| - \frac{c_k}{a_{hk}} ||b_h a_{h1} \dots a_{hn}||,$$

and now because of our assumption  $C4$ , together with the fact that  $-c_k/a_{hk}$  is positive, we conclude from (2.13) that the transformed cost row is lexicographically greater than the initial one; i.e., we have strict lexicographic improvement in the cost row.

However, in case there are ties in (2.8), i.e., if  $h$  is not uniquely defined by (2.8) then we must choose  $h$  so as to preserve condition  $C4$ . This is achieved by the following requirement:

$$(2.14) \quad \frac{1}{a_{hk}} ||b_h a_{h1} \dots a_{hn}|| = \text{lexmin} \left\{ \frac{1}{a_{ik}} ||b_i a_{i1} \dots a_{in}|| \mid a_{ik} > 0 \right\}.$$

This choice of  $h$  is unique since no two rows of  $M$  can be equal.

To see that  $M^*$  now satisfies  $C4$  we first observe that row  $h$  of  $M^*$  is

$$(2.15) \quad \frac{1}{a_{hk}} ||b_h a_{h1} \dots a_{hn}||$$

and is lexicographically positive because the  $h$ -th row of  $M$  was. Next, if  $i = h$  we have for the  $i$ -th row of  $M^*$

$$(2.16) \quad ||b_i^* a_{i1}^* \dots a_{in}^*|| = ||b_i a_{i1} \dots a_{in}|| - \frac{a_{ik}}{a_{hk}} ||b_h a_{h1} \dots a_{hn}|| \\ = a_{ik} \left( \frac{1}{a_{ik}} ||b_i a_{i1} \dots a_{in}|| - \frac{1}{a_{hk}} ||b_h a_{h1} \dots a_{hn}|| \right).$$

Now, if  $a_{1k} \leq 0$  the first line of (2.16) shows that row (i) is either unchanged or is lexicographically increased; and if  $a_{1k} > 0$  our choice of  $h$  via (2.14) guarantees that the difference in the second line of (2.16) is lexicographically negative. Thus  $C_4$  holds for the transformed matrix  $M^*$ . This completes the proof of Theorem 2A.

Lemma 2B. Let  $M$  and  $M^*$  define equivalent linear programming problems (in the sense of Theorem 1B) in canonical form with respect to the same basic sequence  $S$ . Then  $M = M^*$ .

Proof: Suppose that (1.10) holds for  $Q$  given by (1.13). Then using subscripts to denote columns of a matrix we have

$$M_j^* = QM_j ; \quad j = 0, \dots, n .$$

In particular, if  $j \in S$ , say  $j = s_h$  we have  $M_{s_h}^* = M_{s_h} = U_{h+1}$  the  $(h+1)$ st unit vector and so

$$U_{h+1} = QU_{h+1} = Q_{s_h} , \quad h = 1, \dots, p .$$

But  $Q_0 = U_1$ ; hence  $Q = I_{p+1}$  and  $M = M^*$ .

Theorem 2C. (first convergence Theorem) Consider a linear programming problem in canonical form. Then after a finite number of pivot operations one of the terminal states (i) or (ii) of Theorem 2A is attained.

First permute columns, if necessary, so that  $(C_4)$  holds. Then

Proof: Consider a sequence  $M = M_0, M_1, M_2, \dots, M_t$  of canonical tableau matrices all in case (iii) and satisfying  $(C_4)$  and where for each  $i (i = 1, \dots, t)$   $M_i$  is obtained from  $M_{i-1}$  by pivoting on a position  $(h_i, k_i)$  selected so that  $c_{k_i}^{(i-1)}$  is a negative entry in the cost row of  $M_{i-1}$  and  $h_i$  is determined using (2.14). Then, since by the conclusion for case (iii) of Theorem 2A, there is lexicographic improvement in the zero-th row at each stage, no two matrices in the sequence can be equal. It follows from Lemma 2B that no two of the corresponding basic sequences can be equal. Hence,  $t+1$  cannot exceed the number of possible basic sequences. We conclude  $t$  is less than the number  $P(n, p)$  of permutations of  $n$  objects taken  $p$  at a time and therefore  $t$  is bounded. On the other hand, the sequence can be extended unless one of the terminal states (i) or (ii) has been reached. We conclude that a terminal state can be reached after less than  $P(n, p)$  pivots.

( 13 )

Fortunately, in practice the simplex algorithm has been found to terminate after relatively few pivots, although there is as yet not a mathematically established bound which comes close to explaining computational results.

### 3. Reduction to Canonical Form

Having obtained a convergence theorem for problems in canonical form, we now turn to the matter of reduction to canonical form. We organize the question of reduction under sequence convergence theorems, each covering a more general case than the former, until we reach one which applies to the most general form of linear programming.

**Theorem 3A. (Second convergence theorem.)** Let  $M$  represent a problem in standard form for which C1 holds. Then there exists a finite sequence of pivot operations which will terminate in an equivalent problem  $M^*$  which is in state (i) or (ii) of Theorem 2B or which

(iv) has an infeasible constraint of the form "negative scalar equals sum of products of non-negative scalars",

i.e.,  $M^*$  has a row of the form

$$(3.1) \quad ||b_i \ a_{i1} \ \dots \ a_{in}||$$

where  $b_i < 0$  and  $a_{ij} \geq 0 \ j = 1, \dots, n$ .

We sometimes represent states (i), (ii), (iv) symbolically by

(i)

		+
+		

(ii)

		-	
+		-	

(iv)

-	⊕

Proof: Conceptually, we may achieve C2 by using the equations to eliminate the basic variables from the cost function. Analytically, we proceed as follows. Let

$$(3.2) \quad Z^T = ||z_1 \dots z_n|| = C_S^T A, \quad z_j = C_S^T B;$$

then successive pivots on the  $p$  positions,  $(1, s_1), (2, s_2), \dots, (p, s_p)$  will yield a new cost row  $||-d^* \ C^{*T}||$  where

$$(3.3) \quad C^* = C - Z, \quad -d^* = -d - z_0,$$

and will not change any of the last  $p$  rows of  $M$ . Since  $A_{sj}$  is the  $j$ -th unit vector in  $p$ -space,  $z_{sj} = c_{sj}$  so that  $z_{sj}^* = 0$  ( $j = 1, \dots, p$ ); hence, C2 is now satisfied.

Tables 3.1 and 3.2 give, respectively, a schematic representation and a numerical example of this process.

Next, if C3 is not satisfied, we consider an auxiliary subproblem,

$$(3.4) \quad M' = \begin{array}{|c|c|} \hline -d' & C'^T \\ \hline B' & A' \\ \hline \end{array}$$

whose zero-th row is any row of  $M$  for which the constant term is negative, and where  $||B' \ A'||$  consists of all rows of  $M$  for which the constant term is positive or zero. Now  $M'$  is in canonical form, and we apply the first convergence theorem to  $M'$  to obtain a sequence of pivot positions leading to either state (i) or state (ii). We actually pivot in all of  $M$ , so as to preserve C1 or C2 for it. Let  $M^*$  and  $M^{*1}$  represent the resulting problem and auxiliary subproblem.

TABLE 3.1

## CALCULATION OF CANONICAL COSTS

$C_S$	$-d$	$C^T$
	B	A
	$z_0$	$Z^T$
	$-d^*$	$C^{*T}$

$$z_0 = C_S^T B$$

$$z_j = C_S^T A_j \quad (j = 1, \dots, n)$$

$$-d^* = -d - z_0$$

$$c_j^* = c_j - z_j \quad (j = 1, \dots, n)$$

TABLE 3.2

## CALCULATION OF CANONICAL COSTS, ILLUSTRATIVE EXAMPLE

B	$A_1$	$A_2$	$A_3$	$A_4$	$A_5$	$A_6$
3	2	5	4	-3	-2	1
3	2	0	1	4	0	-5
4	-1	0	0	2	1	2
1	3	1	0	-4	0	2

(a) Initial tableau,  $S = (3, 5, 2)$ 

$C_S$	B	$A_1$	$A_2$	$A_3$	$A_4$	$A_5$	$A_6$
3	3	2	5	4	-3	-2	1
3	3	2	0	1	4	0	-5
-2	4	-1	0	0	2	1	2
5	1	3	1	0	-4	0	2
$z_j$	9	25	5	4	-8	-2	-14
$c_j^*$	-6	-23	0	0	5	0	15

(b) Calculation of canonical costs  $c_j^*$ 

B	$A_1$	$A_2$	$A_3$	$A_4$	$A_5$	$A_6$
-6	-23	0	0	5	0	15
3	2	0	1	4	0	-5
4	-1	0	0	2	1	2
1	3	1	0	-4	0	2

(c) Canonical tableau

If state (i) obtains for  $M^{*1}$  and  $-d^{*1} < 0$ , then  $M^*$  is in state (iv) and, therefore, the problem is infeasible. If state (i) obtains for  $M^{*1}$  and  $-d^{*1} \geq 0$ , we form a new subproblem,  $M^{*11}$ , which is in canonical form and has at least one more row than  $M^*$  had. [Actually, if  $-d^{*1} \geq 0$  after any pivot, we need not continue to a terminal state but can immediately pass to a "larger"  $M^{*11}$ .]

If state (ii) obtains with, say,  $c_k^{*1} < 0$  and  $A_k^{*1} \leq 0$ , we choose the cost row of the subproblem as pivot row and column  $k$  as pivot column. Then  $c_k^{*1}$  is the pivot element. We claim that, after pivoting, the resulting  $M^{**}$  will be feasible every row of  $M^{*1}$  (including its cost row) and, hence, we may form a "larger"  $M^{*11}$  as the continuing auxiliary subproblem.

To verify this claim, we observe, first, that  $-d^{**1} = (-d^{*1})/c_k^{*1} > 0$ , since by hypothesis both numerator and denominator are negative. Finally, if  $i$  is any row of  $M^*$ , we have

$$b_i^{**} = b_i^* - a_{ik}^* (-d^{*1})/c_k^{*1} \geq b_i^*,$$

since by hypothesis the first term in the fraction is nonpositive (as an element of  $A_k^{*1}$ ) and the other two are negative.

Thus, after a finite number of auxiliary subproblems of increasing size, we must either reach state (iv) or obtain an  $M^*$  with  $B^* > 0$  and hence, by Theorem 3A, eventually reach state (i) or state (ii) for the full problem. This completes the proof of Theorem 3A.

Remark: If, serendipitously, rows of  $M$  not in  $M^*$  become feasible before the cost row of  $M^*$  does, one can immediately pass to a larger subproblem. The desirability of including the search for this phenomenon in a computer program is still an open question.

Table 3.3 illustrates pivot defined by an auxiliary subproblem. Note that, in this example, one pivot achieves one more feasible row. An additional pivot in position (3,1) will demonstrate infeasibility of the problem.

Theorem 3B. (Third convergence theorem.) Let  $M$  represent a problem in standard form. Then there exists a finite sequence of pivots which will terminate in an equivalent problem  $M^*$ , which either

(v) has a constraint of the form "nonzero scalar equals zero",

TABLE 3.3

## EXAMPLE ILLUSTRATING AUXILIARY PROBLEM

B	A <sub>1</sub>	A <sub>2</sub>	A <sub>3</sub>	A <sub>4</sub>	A <sub>5</sub>	A <sub>6</sub>	A <sub>7</sub>
4	0	0	3	0	2	0	4
3	0	0	3	1	2	0	3
2	1	0	-1	0	2	0	2
-2	0	0	2	0	-3	1	1
-2	0	1	3	0	4	0	-1

(a) Initial tableau

-2	0	0	2	0	-3	1	1
3	0	0	3	1	2	0	3
2	1	0	-1	0	2	0	2

(b) Auxiliary tableau

AFTER PIVOT AT (2, 5)

2	-1	0	$\frac{1}{2}$	0	0	0	0
1	-1	0	2	1	0	0	1
1	$\frac{1}{2}$	0	$-\frac{1}{2}$	0	1	0	1
1	$\frac{3}{2}$	0	$\frac{1}{2}$	0	0	1	4
-6	-2	1	5	0	0	0	-5

i.e.,  $M^*$  has a row of the form

$$(3.5) \quad ||b_i \ a_{i1} \ \dots \ a_{in}||$$

where  $b_i \neq 0$  and  $a_{i1} = \dots = a_{in} = 0$ , or satisfies C1 after deletion of any zero rows.

Proof: Suppose that  $S = (s_1, \dots, s_q)$  is a sequence of  $q$  distinct integers, where  $0 \leq q < p$  such that

$$(3.6) \quad A_{s_1} = I_1, \dots, A_{s_q} = I_q.$$

Then let  $h = q + 1$  and consider the  $h$ -th row of  $M$ . If it satisfies (v), the problem is infeasible and we stop. If the entire row is zero, we delete it and proceed with the resulting tableau. Otherwise, there exists  $k$  such that  $a_{hk} \neq 0$ . Because of (3.6),  $k \notin S$ . We now pivot on position  $(h, k)$  to obtain a new tableau with one more canonical column. We let  $s_{q+1} = k$ ,  $S' = (s_1, \dots, s_{q+1})$ , and (3.6) holds for the transformed tableau relative to  $S'$ . Since each stage either stops with (v) or decreases  $p - q$ , the process cannot continue indefinitely, and the theorem is established.

The first three reduction theorems show that, for any problem in standard form, there exists a finite sequence of pivots, which results in an equivalent problem  $M^*$  in one of the terminal forms (i), (ii), (iv) or (v). We now define the general linear program and show how to reduce it to standard form; thus, we will complete our proof of convergence of the simplex algorithm.

The most general form of linear programming problem has both nonnegative and free variables and both inequalities and equations. Suppose that there are  $n_1$  nonnegative variables,  $n_2$  free variables,  $p_1$  inequalities, and  $p_2$  equalities. The problem takes the form

$$(3.7) \quad \text{Minimize } z = d + C_1^T X_1 + C_2^T X_2$$

subject to the constraints

$$(3.8) \quad A_{11} X_1 + A_{12} X_2 \geq B_1$$

$$(3.9) \quad A_{21} X_1 + A_{22} X_2 = B_2$$

$$(3.10) \quad X_1 \geq 0$$

$$(3.11) \quad X_2 \text{ free}$$

We introduce a vector  $X_3$  with  $p_1$  components, set  $C_3 = 0$ , replace (3.7) and (3.8) by

$$(3.7') \quad \text{Minimize } z = d + C_1^T X_1 + C_2^T X_2 + C_3^T X_3 ,$$

$$(3.8') \quad A_{11} X_1 + A_{12} X_2 - I_{p_1} X_3 = B_1 ,$$

and add the constraint

$$(3.12) \quad X_3 \geq 0 .$$

We call  $X_3$  a slack vector.

The resulting problem would be in standard form were it not for the free variables. A little reflection leads to the observation that basic free variables present no difficulty but that nonbasic (i.e., independent) free variables upset the logic of the simplex method. This suggests that we look for pivot operations which will bring as many free variables as possible into the basis. We will then be able to lay the corresponding equations aside, temporarily, and deal with a smaller problem in standard form.

We proceed inductively and assume for some  $q$  with  $0 \leq q \leq n_2$  that there are  $q$  free variables in the basis and that by permuting rows (if necessary) these free variables are in rows  $1, \dots, q$  and that the basic free variable in row  $i$  is in column  $s_i$  ( $i = 1, \dots, q$ ). If  $q < n_2$  there exists another free column  $k$  and for it there are three cases to consider (set  $p = p_1 + p_2$ )

$$(a) \quad a_{q+1,k} = \dots = a_{p,k} = c_k = 0$$

$$(b) \quad a_{q+1,k} = \dots = a_{pk} = 0, \quad c_k \neq 0$$

$$(c) \quad \text{for some } h \text{ with } q < h \leq p, \quad a_{hk} \neq 0 .$$

If case (a) holds for every nonbasic free column, or if  $q = n_2$ , we consider the subproblem which remains when we delete all of the free columns and the first  $q$  rows. This subproblem is in standard form and therefore can be handled by the methods already developed. If the subproblem has a basic optimal feasible solution  $(X_1^0, X_3^0)$ , we obtain a basic optimal feasible solution  $(X_1^0, X_2^0, X_3^0)$  for the main problem by setting each nonbasic free variable equal to zero and by setting

$$\begin{aligned}
 (3.13) \quad x_{s_i}^0 &= b_i - (a_{i1} x_1^0 + \dots + a_{in_1} x_{n_1}^0) \\
 &\quad - (a_{i,n_1+n_2+1} x_{n_1+n_2+1}^0 + \dots + a_{i,n_1+n_2+n_3} x_{n_1+n_2+n_3}^0) \\
 i &= 1, \dots, q.
 \end{aligned}$$

Moreover, the minimum value of  $z$  is the same for the main problem and the subproblem.

If case (b) holds, then either there is no lower bound for the objective function or the problem is not feasible. For, suppose  $X^0$  is any feasible solution with corresponding objective value  $z^0$ . Then define a vector  $Y$  as follows:

$$\begin{aligned}
 (3.14) \quad y_{s_i} &= -a_{ik} \quad i = 1, \dots, s_q \\
 y_k &= 1 \\
 y_i &= 0 \text{ for all other } i.
 \end{aligned}$$

Then  $Y$  makes the left-hand sides of (3.7') and (3.8) zero, so that the vector

$$X = X^0 + tY$$

is (i) feasible for all real numbers  $t$ , and (ii) yields  $z = z^0 + tc_k$ . It is clear that the constraint equations (3.7') and (3.8) are satisfied by  $X$ . Moreover, for all  $t$ ,  $X_1 = X_1^0 \geq 0$  and  $X_3 = X_3^0 > 0$ , so that  $X$  satisfies all of the feasibility requirements. Next,  $z = c_0 + C^T X = (c_0 + C^T X^0) + C^T Y = z^0 + tc_k$  (since  $c_{s_i} = 0$ ,  $i = 1, \dots, q$ ). Now, by giving  $t$  the sign opposite that of  $c_k$  and by making the absolute value of  $t$  sufficiently large, we can make  $z$  take negative values with arbitrarily large absolute value; i.e., the objective function has no lower bound. Thus, if case (b) occurs, we stop knowing that the problem has no solution.

Finally, if case (c) holds, we (i) pivot at position  $(h,k)$ , (ii) permute row  $h$  and row  $q+1$ , (iii) set  $s_{q+1} = k$  and thereby have completed an inductive step. We then iterate the process until eventually we reach case (a), case (b), or  $q = n_2$ .

The three cases are illustrated, respectively, in Tables 3.4, 3.5, and 3.6. The second tableau in Table 3.6 comes under case (a). In all of the examples  $x_1, x_2$  are nonnegative variables,  $x_3, x_4$  are free

TABLE 3.4

CASE (a) TRANSITION TO SUBPROBLEM

B	A <sub>1</sub>	A <sub>2</sub>	A <sub>3</sub>	A <sub>4</sub>	A <sub>5</sub>
6	0	4	0	0	0
3	2	2	1	2	1
2	0	3	0	0	1
4	1	-2	0	0	0

$$p_1 = 1, p_2 = 2, n_1 = 2, n_2 = 2, q = 1, s_1 = 3, k = 4$$

$$\begin{aligned} \begin{vmatrix} x_1^0 \\ x_3^0 \end{vmatrix} &= \begin{vmatrix} 4 \\ 0 \\ 2 \end{vmatrix}, z^0 = -6 \\ x^0 &= \begin{vmatrix} x_1^0 \\ x_2^0 \\ x_3^0 \end{vmatrix} = \begin{vmatrix} 4 \\ 0 \\ -7 \\ 0 \\ 2 \end{vmatrix} \end{aligned}$$

TABLE 3.5

CASE (b) UNBOUNDED OBJECTIVE FUNCTION

B	A <sub>1</sub>	A <sub>2</sub>	A <sub>3</sub>	A <sub>4</sub>	A <sub>5</sub>
0	1	3	-2	0	4
3	1	2	3	1	1
2	-1	3	0	0	-3
-4	2	-6	0	0	3

$$p_1 = 1, p_2 = 2, n_1 = 2, n_2 = 2, q = 1, s_1 = 4, k = 3$$

$$x^0 = \begin{vmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{vmatrix}, z^0 = 4, y = \begin{vmatrix} 0 \\ 0 \\ 1 \\ -3 \\ 0 \end{vmatrix}$$

$$z = 4 - 2t \rightarrow -\infty \text{ as } t \rightarrow +\infty$$

TABLE 3.6

CASE (c) INCREASE IN  $q$ 

B	$A_1$	$A_2$	free slack		
			$A_3$	$A_4$	$A_5$
0	1	3	-2	0	4
3	1	2	3	1	1
2	-1	3	1	0	-3
-4	2	-6	-3	0	3

 $p_1 = 1, p_2 = 2, n_1 = 2, n_2 = 2, q = 1, s_1 = 4, k = 3, h = 2$ 

4	-1	9	0	0	-2
-3	4	-7	0	1	10
2	-1	3	1	0	-3
2	-1	3	0	0	-6

 $q = 2, s_1 = 4, s_2 = 3$

variables, and  $x_5$  is a slack (nonnegative) variable. The tableau corresponding to the subproblem in the first example is

B	$A_1$	$A_2$	$A_5$
6	0	4	0
2	0	3	1
4	1	-2	0

and is in optimal form with corresponding solution

$$\begin{bmatrix} 4 \\ 0 \\ 2 \end{bmatrix}$$

as listed.

The subproblem obtained from the second tableau in the third example is

B	$A_1$	$A_2$	$A_5$
4	-1	9	-2
2	-1	3	-6

After pivoting in position (1,2) of this subproblem, i.e., position (3,2) of the corresponding full problem, we obtain

-2	2	0	16
2/3	-1/3	1	-2

which is in case (1), so that the original problem has the optimal solution.

$$z = 2 \quad X = \begin{bmatrix} 0 \\ 2/3 \\ 0 \\ 5/3 \\ 0 \end{bmatrix}$$

We summarize these results in our final convergence theorem.

**Theorem 3C.** (Fourth convergence theorem.) Given a general linear programming problem (3.7) through (3.11), there exists a finite sequence of pivot operations which either leads to an optimal solution, demonstrates infeasibility, or demonstrates unbounded cost.

Actually, we have done more, because we have given an algorithm which selects the successive pivot positions and which identifies infeasibility and unboundedness in several specific forms.

There is an alternate method for handling free variables, namely, the replacement of  $X_2$  throughout

$$(3.15) \quad X_2 = X_2' - X_2''$$

where

$$(3.16) \quad X_2' \geq 0, \quad X_2'' \geq 0;$$

we thus have an equivalent problem with only nonnegative variables. This alternate method is, for theoretical purposes, quite satisfactory; however, it fails to take any advantage of the flexibility which free variables afford via the reduction to a smaller problem after the free variables are pivoted into the basis.

## ERRATA

- 7.6          Replace  $(1, \dots, s)$  by  $(1, \dots, n)$
- 8.11        Replace  $is$  by  $in$
- 10 Top figure. Replace  $b_y$  by  $b_j$
- 11.3        Replace  $i = h$  by  $i \neq h$
- 12.4        Replace negative by positive
- 19.9        Delete comma
- 14.4        Should read ...under a sequence of convergence.
- 26 (4.2)    Replace  $d^*$  by  $-d^*$
- (4.4)    Replace  $a_{1k}$  by  $c_k$
- Replace  $a_{2k}$  by  $a_{1k}$
- Replace  $c_k$  by  $a_{pk}$
- 31.7        Replace  $M$  by  $\Sigma$
- 36.6        Replace  $X_{22}$  by  $X_2$
- 37.12       Replace  $C_k^*$  by  $c_k^*$  twice
- 41.5        Should read ... the protein illustration  $b'_h$  represents the...
- 41.11       Replace second  $\tau$  by  $t$
- 41.20       Replace  $\tau$  by  $t$

INTRODUCTION TO LINEAR PROGRAMMING

R. M. Thrall

July 1967

# INTRODUCTION TO LINEAR PROGRAMMING

R. M. Thrall

<u>Chapter</u>	<u>Page</u>
1. Introduction .....	1
2. Canonical form, Convergence .....	6
3. Reduction to Canonical Form .....	14
4. A Computational Check; The Revised Simplex Method ....	26
5. Duality .....	30
6. Interpretation .....	39

4. A Computational Check; The Revised Simplex Method

Let

$$(4.1) \quad M = \left\| \begin{array}{cc} -d & C^T \\ B & A \end{array} \right\|$$

be the initial tableau matrix for a problem in standard form, and let

$$(4.2) \quad M^* = \left\| \begin{array}{cc} d^* & C^{*T} \\ B^* & A^* \end{array} \right\|$$

be a second tableau matrix for the same problem which we assume to be in canonical form relative to some basic sequence  $S^* = (s_1^*, \dots, s_p^*)$ .

For the case of a full tableau matrix  $M$  and any basic sequence  $S = (s_1, \dots, s_p)$  we modify our previous notation for submatrices slightly and thus write

$$(4.3) \quad M_S = \left\| \begin{array}{cc} 1 & C_S^T \\ 0 & A_S \end{array} \right\|$$

to designate the square matrix of degree  $p + 1$  consisting of the first unit vector of  $V_{p+1}$  followed by columns  $s_1, \dots, s_p$  of  $M$  (we consider  $\left\| \begin{array}{c} -d \\ B \end{array} \right\|$  to be column 0 of  $M$ ).

The process of pivoting on position  $(h,k)$  of  $M$  can be regarded as premultiplication of  $M$  by the matrix

$$(4.4) \quad F = \left\| \begin{array}{cccccc} 1 & 0 & \dots & -a_{1k}/a_{hk} & \dots & 0 \\ 0 & 1 & \dots & -a_{2k}/a_{hk} & \dots & 0 \\ \vdots & \vdots & & \vdots & & \vdots \\ \vdots & \vdots & & 1/a_{hk} & & \vdots \\ \vdots & \vdots & & \vdots & & \vdots \\ 0 & 0 & \dots & -c_k/a_{hk} & \dots & 1 \end{array} \right\|.$$

Here  $F$  differs from the identity matrix  $I$  only in column  $h + 1$  which is indicated in (4.4), or in terms of  $M$ ,  $f_{ih}^{+1} = -m_{ik}/m_{hk}$  for all  $i \neq h$  and  $f_{hh}^{+1} = 1/m_{hk}$ . It follows that a sequence of pivot operations can be effected by a single matrix multiplication, and the reverse sequence can be achieved by multiplying by the inverse of this same matrix. Suppose then that

$$(4.5) \quad M^* = QM, \quad M = Q^{-1} M^*$$

Since we never pivot in the initial row or column, each pivot matrix  $F$  has initial column

$$\begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

inverse  $Q^{-1}$ .

Now

$$(4.6) \quad M_{S^*}^* = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} M_{s_1}^* \dots M_{s_p}^* = I_p + 1$$

It follows from (4.5) that for each  $j$ ,  $M_j = Q^{-1} M_j^*$ ; also  $\begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$ ;

hence we have

$$(4.7) \quad M_{S^*} = Q^{-1} M_{S^*}^* = Q^{-1}, \quad \text{or}$$

$$(4.8) \quad M = M_{S^*} M^*.$$

This equation provides a numerical check of the accuracy of  $M^*$  under the hypothesis that  $M^*$ , but not necessarily  $M$ , is in canonical form. We call it a reverse check since it goes from  $M^*$  to  $M$  by premultiplication with a "submatrix" of  $M$ .

If  $M$  is in canonical form with respect to a basic sequence  $S = (s_1, \dots, s_p)$ , then

$$(4.9) \quad M^* = M_{S^*}^* M, \quad \text{i.e., } Q = M_{S^*}^*.$$

We call (4.9) a direct check since it multiplies  $M$  by a "submatrix" of  $M^*$  to obtain  $M^*$ . If a check (either direct or reverse) discloses an error, one can then check in intermediate stages and soon locate the source of the error. For example, if  $M^*$  is the result of 12 cycles, first check at Cycle 6. If this is correct, check at Cycle 9; if it is incorrect check at Cycle 3, etc.

We observe that when both  $M$  and  $M^*$  are in canonical form, then

$$(4.10) \quad M_{S^*}^* M_S^* = I_{p+1};$$

thus we may regard the simplex method as being a systematic procedure for calculating inverses of the "submatrices"  $M_{S^*}^*$  of  $M$  and for specifying, by selection of  $S^*$ , which submatrices to invert.

In applying (4.5), given  $M$  and  $Q$ , there is no need to calculate all of  $M^*$ . Indeed, if we first calculate the initial row of  $M^*$  we can either (i) find an index  $k$  for which  $c_{hk}^* < 0$  or (ii) conclude the  $M^*$  is in optimal form.

In case (i) we calculate column  $k$  of  $M^*$  and conclude either (iii) there is some  $a_{ik} > 0$  or (iv) the problem has no optimal solution. In case (iii) we calculate column 0 (the  $B$  column) and apply the pivot selection rule to determine the next pivot position  $(h,k)$ . Let  $M'$  be the matrix obtained from  $M^*$  after pivoting at  $(h,k)$ . Then

$$(4.11) \quad M' = F^* M^*$$

where  $F^*$  is given by (4.4) (calculated from  $M^*$  rather than from  $M$ ). Then, from (4.5) we get

$$(4.12) \quad M' = Q' M \text{ where } Q' = F^* Q.$$

This completes a full cycle of the algorithm. Note that this form of the simplex algorithm requires (i) keeping a permanent record of the initial tableau matrix  $M$ , (ii) keeping a temporary record of the transforming matrix  $Q$  which is discarded in favor of its successor  $Q'$  when it is calculated, and (iii) calculating the final row, the initial column, and one other column of  $M^*$ . These are used to decide the state of  $M^*$  and to determine  $F^*$  in case  $M^*$  is not in a terminal state and then are discarded. One might also wish to (iv) keep a record of the basic sequence  $S'$ .

[The revised simplex method is not recommended for hand calculation. For that matter, hand calculation in any form is clearly hopelessly inefficient for all but very tiny problems. Thus we must evaluate the revised simplex method with machine calculation in mind.]

It is not particularly difficult to devise a computer program which will put into effect the four parts (i) - (iv) of the revised simplex method. This method has several important advantages for digital calculation:

(a) All calculations are based on the initial data, and the accuracy of  $Q$  can be tested by reference to (4.7), i.e.,  $Q$  is the inverse of  $M_S^*$ ;

(b) If the original data is sparse (i.e., has many zero coefficients) the calculations will be simpler and quicker at every cycle since this sparse matrix  $M$  is used throughout. In the ordinary algorithm sparseness is rapidly lost. Moreover, very large sparse problems can be handled with a computer memory which could not touch nonsparse problems of the same size;

(c) The fact that not all of  $M^*$  need be calculated gives substantial computing economy.

In most current computer programs the method, as described, is modified in several ways, of which we discuss only one. Instead of storing  $Q$  we store only each factor  $F'$ , and for  $F'$  we store only its non-identity column together with a record of the pivot position. The algorithm in this form is called the "product form of the revised simplex method" and several computer algorithms for this method are in current use. We do not discuss the details here.

5. Duality

We consider a pair of related problems

$$(5.1) \quad DV \leq R, \quad V \geq 0, \quad \text{maximize } P^T V$$

and

$$(5.2) \quad D^T U \geq P, \quad U \geq 0, \quad \text{minimize } R^T U,$$

where  $D$  is a  $p$ -by- $n$  matrix and we assume that problem (5.1) has an optimal feasible solution.

The introduction of slack variables  $Y$  in (5.1) leads to a problem in standard form

$$(5.3) \quad AX = B, \quad X \geq 0, \quad \text{minimize } z = d + C^T X$$

where

$$(5.4) \quad A = \begin{bmatrix} D & I \end{bmatrix}, \quad X = \begin{bmatrix} V \\ Y \end{bmatrix}, \quad C = \begin{bmatrix} -P \\ 0 \end{bmatrix}, \quad B = R, \quad d = 0,$$

and  $C1$  and  $C2$  hold for the basic sequence  $S = (m+1, \dots, m+p)$ .

Thus as initial tableau we have

$$(5.5) \quad M = \begin{bmatrix} -d & C^T \\ B & A \end{bmatrix} = \begin{bmatrix} 0 & -P^T & 0 \\ R & D & I_p \end{bmatrix};$$

let the final (optimal) tableau be

$$(5.6) \quad M^* = \begin{bmatrix} -d^* & C^{*T} \\ B^* & A^* \end{bmatrix} = \begin{bmatrix} -d^* & C_G^{*T} & C_S^{*T} \\ B^* & A_G^* & A_S^* \end{bmatrix}.$$

Next, we let

$$(5.7) \quad U_0 = C_J^*,$$

and then from (4.9) we get

$$(5.8) \quad -d^* = U_0^T R, \quad C_G^{*T} = -P^T + U_0^T D.$$

(31)

Now, since (5.6) is in optimal form, we have

$$(5.9) \quad C_S^* \geq 0, \quad C_G^* \geq 0,$$

or, equivalently,

$$(5.10) \quad U_0 \geq 0, \quad D^T U_0 \geq P.$$

Hence,  $U_0$  is a feasible vector for the dual problem.

Let  $X_0 = \begin{pmatrix} V_0 \\ Y_0 \end{pmatrix}$  be the optimal vector for the primal problem. Then

$$(5.11) \quad d^* = C^T X_0 = -P^T V_0;$$

hence

$$(5.12) \quad P^T V_0 = R^T U_0 = -d^*.$$

For any feasible vectors  $V$  and  $U$  for (5.1) and (5.2) we have

$$(5.13) \quad R^T U = U^T R \geq U^T D V \geq P^T V.$$

From (5.13), (5.12) and the feasibility of  $V_0, U_0$ , it follows that both  $V_0$  and  $U_0$  are optimal. Since  $U_0$  appears as part of the initial row of  $M^*$ , we see that the simplex algorithm solves both of the related problems simultaneously and that the objective values are equal.

The relationship between the problems (5.1) and (5.2) is an interesting special instance of duality in linear programming. Even this special instance has a number of important applications; for example, the minimax theorem of game theory is a consequence of (5.11). However, the special form of (5.1) leads us to ask how to define duality for more general linear programming problems. Indeed, we might ask about the characteristics of a duality relation in still more general situations.

Let  $\Sigma$  be a set of objects. A mapping  $T$  of  $\Sigma$  into itself is said to be involutory if  $\forall a \in \Sigma, T(T(a)) = a$ . If  $\sim$  is an equivalence relation on  $\Sigma$ ,  $T$  is said to be equivalence preserving if  $a \sim a'$  implies  $T(a) \sim T(a')$ .

We say that  $T$  is meaningful relative to some relation  $R$  on  $\Sigma$  if for all  $a \in \Sigma, a R T(a)$ .

For example, let  $\Sigma$  be the set of all linear programming problems each of which has either form (5.1) or the form

$$(5.2') \quad F U \geq Q, \quad U \geq 0, \quad \text{minimize } S^T U$$

and let  $T$  be the mapping used above which sends the problem of (5.1) into (5.2) and conversely. In more detail, if we let  $\Sigma_1$  denote the set of all problems of form (5.1), let  $\Sigma_2$  denote the set of all problems of form (5.2'), and let  $\Sigma = \Sigma_1 \cup \Sigma_2$  then  $T$  interchanges the two subsets  $\Sigma_1, \Sigma_2$  and when applied twice to any problem in  $\Sigma$  gives back that same problem. The involutory property of  $T$  depends on the fact that transposing any matrix twice gives back that same matrix. This mapping  $T$  is not only involutory but also is meaningful relative to the relation

$$(5.14) \quad \text{optimal value } a = \text{optimal value } T(a).$$

More generally, a mapping  $T$  is meaningful if there are useful theorems connecting  $a$  and  $T(a)$ ; the duality theorems of linear programming are of this nature.

A mapping  $T$  which has all three of these properties i.e. which is involutory, equivalence preserving, and meaningful qualifies as the very "best" type of duality. We obtain such a duality for linear programming under a rather general equivalence relation.

We now denote by  $\Sigma$  the set of all linear programming problems in the general form (3.7) - (3.11). Given any such problem called the primal problem, we show in Table (5.1) how to define a new problem in  $\Sigma$  called the dual problem related to the given primal. We may characterize the primal problem by the  $(p_1 + p_2 + 1)$  -by-  $(n_1 + n_2 + 1)$  matrix

$$(5.14) \quad G = \begin{vmatrix} -d & c_1^T & c_2^T \\ B_1 & A_{11} & A_{12} \\ B_2 & A_{21} & A_{22} \end{vmatrix}$$

together with the ordered set of integers  $(p_1, p_2, n_1, n_2)$  which indicates the nature of each constraint and variable. Then the dual problem is characterized by the matrix  $-G^T$  together with the sequence  $(n_1, n_2, p_1, p_2)$ . Clearly, the correspondence between primal and dual is involutory, since, in particular,  $-(-G^T)^T = G$ . Note that the dual is also a minimization problem; each inequality in the primal system corresponds to a non-negative variable in the dual; each equation in the primal system corresponds to a free variable in the dual. Corresponding to each non-negative variable in the primal system is an inequality whose coefficients are the negatives of the coefficients of this variable and whose right hand side is the negative of the coefficient of the variable in the primal objective function;

Primal	Dual
$\min C_1^T X_1 + C_2^T X_2 + d = z$	$\min -B_1^T U_1 - B_2^T U_2 - d = z'$
$p_1: A_{11} X_1 + A_{12} X_2 \geq B_1$	$U_1 \geq 0$
$p_2: A_{21} X_1 + A_{22} X_2 = B_2$	$U_2 \text{ free}$
$n_1: X_1 \geq 0$	$-A_{11}^T U_1 - A_{21}^T U_2 \geq -C_1$
$n_2: X_2 \text{ free}$	$-A_{12}^T U_1 - A_{22}^T U_2 = -C_2$

Table 5.1. Dual Systems, General Case.

each free variable in the primal corresponds to a similarly defined equation in the dual system. The dual objective function has as coefficients the negatives of the right hand sides of the primal restraints. (The minus sign attached to  $d$  in (5.14) is necessary for uniformity in reading relations from the matrix; the initial row represents the equation  $z-d = C_1^T X_1 + C_2^T X_2$ ).

This duality includes the previous one as a special case if we observe that (5.1) is equivalent to

$$(5.2'') \quad -D V \geq -R, \quad V \geq 0, \quad \text{minimize } -P^T V.$$

Two special cases of the general linear programming problems are of frequent occurrence. The case  $p_2 = q_2 = 0$  (cf. 5.1) is called the symmetric form and is the one that arises naturally in many applications. The case  $p_1 = q_2 = 0$  is called the standard form (cf. Section 1 above). We recall that reduction to standard form is a first step in preparing the problem for application of the simplex algorithm. The dual of a symmetric problem is again symmetric; however, the dual of a problem in standard form is not in standard form.

We saw in Section 3 (cf. (3.7'), (3.8'), and (3.12)) how to change from a problem in symmetric form to one in standard form via introduction of slack variables. Thus, if we can show that a problem in general form is equivalent to one in symmetric form, we will have established that each of the three forms actually includes all problems. To do this we first replace the free vector  $X_2$  by the difference  $X_2' - X_2''$  of two non-negative vectors, and then replace each equation by two oppositely directed inequalities. This yields the symmetric system

$$(5.15) \quad AX \geq P, \quad X \geq 0, \quad \min z = d + C^T X$$

(34)

where

$$A = \begin{vmatrix} A_{11} & A_{12} & -A_{12} \\ A_{21} & A_{22} & -A_{22} \\ -A_{21} & -A_{22} & A_{22} \end{vmatrix},$$

$$B = \begin{vmatrix} B_1 \\ B_2 \\ -B_2 \end{vmatrix}, \quad C = \begin{vmatrix} C_1 \\ C_2 \\ -C_2 \end{vmatrix}, \quad X = \begin{vmatrix} X_1 \\ X'_2 \\ X''_2 \end{vmatrix}.$$

It is easy to verify that the dual of (5.15) is equivalent to the original dual; i.e., our duality is equivalence preserving.

We might ask more precisely in what sense the general primal form is equivalent to (5.15). Clearly, any feasible solution  $X$  of (5.15) defines a feasible solution  $X_1, X_2 = X'_2 - X''_2$  of the original problem and with the same  $z$  value. Conversely, if  $X_1, X_2$  is a feasible solution for the original problem and if  $-t$  is smaller than any component of  $X_2$  then  $X'_2 = X_2 + t E_{p2}, X''_2 = t E_{p2}$  (where for any  $h$ ,  $E_h$  is the vector with  $h$  components all equal to one) gives a feasible solution  $X$  for (5.15) with the same  $z$  value. Hence either the two problems have the same optimal value or neither has an optimal solution. Alternatively, given  $X_2$  one can find unique  $X'_2, X''_2$  with  $X_2^T X''_2 = 0, X'_2 \geq 0, X''_2 \geq 0$ , and  $X_2 = X'_2 - X''_2$ . It can be shown, moreover, that the extreme solutions of the two systems<sup>2</sup> can be put into one-to-one correspondence; this means that the arguments which justify the simplex method can be extended to show how to obtain all extreme solutions of the original problem from the basic solutions of the equivalent problem in standard form.

(35)

We give an illustrative example in which  $p_1 = 1$ ,  $p_2 = 2$ ,  $q_1 = 3$ ,  $q_3 = 2$

Primal	Dual
$\min 2x_1 + 3x_2 + 2x_3 - 3x_4 + 7x_5 + 2 = z$	$\min -3u_1 + 7u_2 - 2u_3 - 2 = z'$
$2x_1 + 3x_2 - 7x_3 - 4x_4 + x_5 \geq 3$	$u_1 \geq 0$
$3x_1 - x_2 - 3x_3 + 5x_4 + 2x_5 = -7$	$u_2 \text{ free}$
$4x_1 + 2x_2 - 6x_3 - 8x_4 - 4x_5 = 2$	$u_3 \text{ free}$
(5.16) $x_1 \geq 0$	$-2u_1 - 3u_2 - 4u_3 \geq -2$
$x_2 \geq 0$	$-3u_1 + u_2 - 2u_3 \geq -3$
$x_3 \geq 0$	$7u_1 + 3u_2 + 6u_3 \geq -2$
$x_4 \text{ free}$	$4u_1 - 5u_2 + 8u_3 = 3$
$x_5 \text{ free}$	$-u_1 - 2u_2 + 4u_3 = -7$

The corresponding primal problem in symmetric form has (cf. (5.16)).

$$A = \begin{bmatrix} 2 & 3 & -7 & -4 & 1 & 4 & -1 \\ 3 & -1 & -3 & 5 & 2 & -5 & -2 \\ 4 & 2 & -6 & -8 & -4 & 8 & 4 \\ -3 & 1 & 3 & -5 & -2 & 5 & 2 \\ -4 & -2 & 6 & 8 & 4 & -8 & -4 \end{bmatrix}, \quad B = \begin{bmatrix} 3 \\ -7 \\ 2 \\ 7 \\ -2 \end{bmatrix}, \quad C = \begin{bmatrix} 2 \\ 3 \\ 2 \\ -3 \\ 7 \\ 3 \\ 7 \end{bmatrix}.$$

We have shown that our duality is involutory and equivalence preserving so far as transitions from general to symmetric to standards forms are concerned. We conclude the present section by stating and proving the general duality theorem for linear programming thus showing that duality is meaningful.

We recall that for a general linear programming problem there are three mutually exclusive possibilities:

(36)

- (a) There is a basic optimal feasible vector,
- (b) The problem is feasible but objective function has no finite lower bound for feasible vectors,
- (c) There is no feasible vector.

The fundamental duality theorem states

- (i) If either primal or dual has an optimal solution so does the other,
- (ii) If both primal and dual are feasible then both have optimal solutions.
- (iii) If either primal or dual is feasible but with no lower bound for the objective function then the other is infeasible.
- (iv) Both primal and dual may be infeasible.
- (v) If (i) holds then

$$(5.18) \quad z_{\text{opt}} + z'_{\text{opt}} = 0$$

(vi) If  $X_1, X_2$ , and  $U_1, U_2$  are any feasible solutions to the primal and dual, respectively, then

$$(5.19) \quad z + z' \geq 0.$$

We first establish (5.19). We have

$$\begin{aligned} z + z' &= (C_1^T X_1 + C_2^T X_2 + d) - (B_1^T U_1 + B_2^T U_2 + d) \\ &\geq C_1^T X_1 + C_2^T X_2 - U_1^T (A_{11} X_1 + A_{12} X_2) \\ &\quad - U_2^T (A_{21} X_1 + A_{22} X_2) \\ &= (C_1^T - U_1^T A_{11} - U_2^T A_{21}) X_1 \\ &\quad + (C_2^T - U_1^T A_{12} - U_2^T A_{22}) X_2 \\ &\geq 0 \end{aligned}$$

Since the first term is a sum of products of non-negative factors and the second is 0. Note that we used all of the feasibility conditions for both primal and dual in establishing this chain of inequalities.

Next, we assume that the primal problem has an optimal solution.

(37)

More precisely, let

$$M = \begin{bmatrix} -d & C_1^T & C_2^T & 0 \\ B_1 & A_{11} & A_{12} & -I_{p_1} \\ B_2 & A_{21} & A_{22} & 0 \end{bmatrix}$$

be the initial matrix, and suppose that

$$M^* = \begin{bmatrix} -d^* & C_1^{*T} & C_2^{*T} & C_3^{*T} \\ B_1^* & A_{11}^* & A_{12}^* & A_{13}^* \\ B_2^* & A_{21}^* & A_{22}^* & A_{23}^* \end{bmatrix}$$

is the optimal tableau. The optimality conditions are  $C_1^{*T} \geq 0$ ,  $C_2^{*T} = 0$ ,  $C_3^{*T} \geq 0$ . The equation  $C_2^{*T} = 0$  is a consequence of the conclusion concerning case (b) (of that section) proved in the argument following (3.14). Consider any free variable  $x_k$ . If it is basic in  $M^*$  then  $C_k^* = 0$  because of (C2) and if it is not basic  $C_k^* = 0$  because of our hypothesis of optimality.

Next let  $Q^{-1} = M_{1*}^{-1}$  be the matrix for which  $M = Q^{-1} M^*$ ; partition  $Q$  by subdivisions the same as for  $M$  giving, say,

$$Q = \begin{bmatrix} 1 & Q_{12} & Q_{13} \\ 0 & Q_{22} & Q_{23} \\ 0 & Q_{32} & Q_{33} \end{bmatrix}$$

and set  $U_1^T = -Q_{12}^T$ ,  $U_2^T = -Q_{13}^T$ . Then, from  $QM = M^*$  and the optimality of  $M^*$  we get

$$\begin{aligned} -d - U_1^T B_1 - U_2^T B_2 &= -d^* \\ C_1^T - U_1^T A_{11} - U_2^T A_{21} &= C_1^{*T} \geq 0 \\ C_1^T - U_1^T A_{12} - U_2^T A_{22} &= C_2^{*T} = 0 \\ U_1^T &= C_3^{*T} \geq 0 \end{aligned} \quad (5.20)$$

(38)

whence it follows that  $U_1, U_2$  is a feasible vector for the dual with objective value  $z' = -d^* = -z_{opt}$  and hence

$$z_{opt} + z' = 0$$

This together with (5.19) shows that  $U_1, U_2$  is optimal for the dual with  $z'_{opt} = z'$  so that (5.18) holds. Since duality is involutory this establishes (i) and (v).

Next, we observe that if both problems are feasible then (5.19) shows that neither can fall in case (b) and this establishes (ii). Finally, (iii) and (iv) together are merely the contrapositive statement of (i) and (ii). We give examples to illustrate possibilities (iii) and (iv). The self dual problem

$$(5.21) \quad \min z = -x_1 \text{ subject to the constraints } 0x_1 \geq 1, \quad x_1 \geq 0$$

illustrates (iv), and the problem

$$(5.22) \quad \min z = -x_1 \text{ subject to the constraints } x_1 \geq 1, \quad x_1 \geq 0$$

illustrates (iii).

6. Interpretation of Final Tableau

Suppose that the  $M$  of (4.1) is the final (optimal) tableau matrix for a linear programming problem. Each number in  $M$  has an interpretation in the language of activity analysis. The entire first column, of course, gives the optimal vector and minimal value of the objective function. Each other column can be used to describe how to adjust the optimal program as a prescribed amount of a nonbasic activity is introduced.

The optimal solution is, of course,

$$(6.1) \quad \begin{aligned} x_{s_i}^S &= b_i \quad (i = 1, \dots, p) \\ x_j^S &= 0 \quad j \notin S \\ z^S &= d \end{aligned}$$

Here  $S = (s_1, \dots, s_p)$  is the optimal basic sequence and as before we denote by  $G$  the set of  $n-p$  indices not in  $S$ . Let  $k$  be a nonbasic index and suppose that we wish to introduce  $t$  units of activity  $k$  into our program. Then the new "best" solution is

$$(6.2) \quad \begin{aligned} x'_{s_i} &= b_i - a_{ik} t \\ x'_j &= 0 \quad j \in G, \quad j \neq k \\ x'_k &= t \\ z' &= d + c_k t \end{aligned}$$

We consider two cases: (i)  $k$  is a slack index and (ii)  $k$  is a natural index. Case (ii) is simpler and we treat it first.

Natural activities, e.g. building picture frames, eating steak, are irreversible hence we restrict  $t$  to nonnegative values. The upper bound for  $t$  is determined by the fact that if any  $a_{ik}$  is positive then  $t$  cannot exceed  $b_i/a_{ik}$  lest  $x'_{s_i}$  become negative. Thus, for  $t$  we have

$$(6.3) \quad 0 \leq t \leq \min \{b_i/a_{ik} \mid a_{ik} > 0; \quad i = 1, \dots, p\}$$

If no  $a_{ik}$  is positive there is no finite upper bound for  $t$ ; occurrence of this in a practical problem is an indication that some constraint may have been overlooked in setting up the model.

A positive value for a slack variable indicates that not all of some resource has been utilized or that some requirement has been more than met. A negative value for a slack variable represents an infeasible solution to the original problem but may have a useful interpretation for slightly modified problems. For example, to consider what would happen if additional units of some resource became available we consider negative values for the corresponding slack variable. Thus, if  $x_k$  measures labor units then  $x_k = -2$  would literally mean a program which exceeded original labor availability by two units and thus (6.2) with  $t = -2$  can be used to describe how the optimal program should be altered to take advantage of two added units of labor. If  $x_k$  measures excess protein then  $t = -2$  would indicate how to best adjust an optimal diet to account for a decision to reduce the minimum daily protein requirement by 2 units.

For a slack index  $k$  the permissible domain for  $t$  is given by

$$(6.4) \quad \max\{b_i/a_{ik} | a_{ik} < 0; \quad i = 1, \dots, p\} \leq t \leq \min\{b_i/a_{ik} | a_{ik} > 0; \\ i = 1, \dots, p\}.$$

Here as in (6.3) we may have one or both limits infinite. For any  $t$  in this interval, (6.2) gives the best adjusted program.

In both (6.3) and (6.4) if  $t$  reaches its extreme value we have the same result as would be given by a pivotal transformation which brings  $k$  into the basic sequence. Any further changes in  $x_k$  must follow the rule which we develop below for changes in basic variables.

Before discussing changes in basic variables we consider the effect of changes in an initial resource or requirement whose corresponding slack variable is basic.

If a basic variable  $x_{sh}$  measures the slack in resource  $h$  then, clearly, no increase in the initial supply of this resource can effect the optimal program or its cost, nor will a decrease by  $t$  units provided that

$$(6.5) \quad t \leq x_{sh}^S$$

Similar reasoning applies to requirements. For example, if the  $h$ -th constraint is the minimum protein requirement, then an increase by  $t$  in this requirement will have no effect if (6.4) holds.

The final tableau matrix does not indicate the amount of the  $h$ -th resource that is used in the program. Let  $M'$  denote the initial tableaux matrix. Then  $b'_h$  is the amount of resource  $h$  that was available initially and  $b'_h - x_{sh}^S$  is the amount used in the optimal program. Similarly, for the vitamin illustration  $b'_h$  was the original minimum daily requirement and  $b'_h + x_{sh}^S$  is the amount of protein actually supplied by the optimal diet.

Next, let  $s_h$  be any basic index and consider the effect of changing  $x_{sh}^S$  by some positive or negative amount  $\tau$ . According to (6.2) this can be done by choosing some nonbasic index  $k$  and letting  $\tau = -a_{hk} \tau$ . In this analysis we wish to preserve the original constraints and hence require  $t \geq 0$ .

For  $\tau > 0$  we must choose  $k$  such that  $a_{hk} < 0$ , then the unit cost for increasing  $x_{sh}$  is  $-c_k/a_{hk}$  and hence the best  $k$  to introduce is given by

$$(6.6) \quad -c_k/a_{hk} = \min \{-c_j/a_{hj} | a_{hj} < 0; j \in G\}$$

Similarly, for  $\tau$  negative the unit cost and best  $k$  are given by

$$(6.7) \quad c_k/a_{hk} = \min \{c_j/a_{hj} | a_{hj} > 0; j \in G\}$$

Whether  $\tau$  is positive or negative when  $t$  takes its upper bound as given by (6.3) say  $\tau = b_s/a_{sk}$  we have for the total increase in cost

$$(6.8) \quad |c_k b_s/a_{sk}|$$

and for further change in  $x_{sh}$  we repeat the entire process beginning with the tableau  $M^*$  obtained from  $M$  after pivoting at  $(s,k)$ . We must take into account the fact that  $M^*$  is not in optimal form but this does not effect (6.2). The fact that one or more of the  $c_j^*$  may be negative gives us no difficulty, provided that we add the requirement  $c_k > 0$  in selecting the new  $k$  in (6.6) and (6.7).

We have discussed the effects of changing the original constant terms  $b_i'$ . Changes in a coefficient  $a_{ik}$  for  $k$  nonbasic has no effect on the optimal program unless  $c_k$  becomes negative when we multiply  $M'$  by  $Q$  to get  $M$ . Of course, the  $a_{ik}$  in (6.2) will have to be adjusted, and if  $c_k$  becomes negative further pivots, beginning in column  $k$  will be required to obtain the new optimal program. Changes in  $a_{ik}$  for  $k$  basic require adjustments in  $Q$  and hence possibly in all of  $M$ ; small changes may not change the optimal basic sequence.

(42)

Changes in  $c'_k$  for  $k$  nonbasic are reflected by exactly the same changes in  $c_k$ ; i.e. if  $c'_k$  is replaced by  $c'_k + t$  then  $c_k$  is also replaced by  $c_k + t$ . The value  $t = -\lambda_k$  is the breakeven value and represents the amount by which  $c'_k$  must be changed to make the  $k$ -th activity competitive economically with the basic activities.

If  $c_{sh}$  is replaced by  $c_{sh} + t$  and we subtract  $t$  times row  $h$  of  $M$  from row zero to restore canonical form  $c_k$  is replaced by  $c_k - t a_{hk}$  for each  $k \in G$ . Hence, if  $t$  lies in the interval

$$(6.9) \quad \max \{c_k/a_{hk} \mid a_{hk} < 0, k \in G\} < t < \min \{c_k/a_{hk} \mid a_{hk} > 0, k \in G\}$$

the optimal basic sequence  $S$  and the corresponding optimal solution (6.1) will remain unchanged except that  $z^S$  will be replaced by  $z^S - t b_h$ .

**NECESSARY CONDITIONS OF OPTIMALITY  
IN CONTROL AND PROGRAMMING**

by

**E. POLAK**

at the

**American Mathematical Society Summer Seminar**

on the

**Mathematics of the Decision Sciences**

**Stanford University**

**July - August 1967**

## NECESSARY CONDITIONS OF OPTIMACITY IN CONTROL AND PROGRAMMING

E. POLAK, (UNIVERSITY OF CALIFORNIA, BERKELEY)

### I. FINITE DIMENSIONAL PROBLEMS

#### Statement of the Basic Problem

Let  $f: E^n \rightarrow E^1$  and  $r: E^n \rightarrow E^m$  be continuously differentiable functions, and let  $\Omega \subset E^n$  be a subset of  $E^n$ . The Basic Problem can be stated as follows:

Find a vector  $\hat{z} \in E^n$  such that

- (i)  $\hat{z} \in \Omega$ ,  $r(\hat{z}) = 0$ ,
- (ii) for all  $z \in \Omega$  with  $r(z) = 0$ ,  $f(\hat{z}) \leq f(z)$ .

We shall call a vector  $\hat{z}$  satisfying (i) and (ii) an optimal solution to the Basic Problem.

#### Necessary Condition for Optimality

The necessary condition to be derived will be stated in the form of an inequality which is valid for all  $\delta z = (\delta z_1, \delta z_2, \dots, \delta z_n)$  in a convex cone "approximation" or "linearization" of the set  $\Omega$ . We shall make use of two kinds of "linearizations" of the set  $\Omega$  at a point  $z$ . The first one will be defined now; the second one will be defined after the proof of Theorem 1, to obtain an extension.

Definition. A convex cone\*  $C(z, \Omega) \subset E^n$  will be called a linearisation of the first kind of the constraint set  $\Omega$  at  $z$  if for any finite collection  $\{\delta z^1, \delta z^2, \dots, \delta z^k\}$  of linearly independent vectors in  $C(z, \Omega)$  there exists an  $\epsilon > 0$ , possibly depending on  $z, \delta z^1, \delta z^2, \dots, \delta z^k$ , such that  $\text{co}\{z, z + \epsilon \delta z^1, \dots, z + \epsilon \delta z^k\}^\dagger \subset \Omega$ .

If the cone  $C(z, \Omega)$  is a linearization of the first kind, then for every  $\delta z \in C(z, \Omega)$  there exists an  $\epsilon_1 > 0$  such that  $z + \epsilon \delta z \in \Omega$  for all  $\epsilon$  such that  $0 \leq \epsilon \leq \epsilon_1$ . The largest cone having this property is given a special name.

Definition. The radial cone to the set  $\Omega$  at a point  $z \in \Omega$  will be denoted by  $RC(z, \Omega)$  and is defined by

$$RC(z, \Omega) = \{\delta z : z + \epsilon \delta z \in \Omega \text{ for all } \epsilon \text{ such that } 0 \leq \epsilon \leq \epsilon_1(z, \delta z) > 0\}$$

\* A set  $C$  is a cone with vertex  $x_0$  if for every  $x \in C, x \neq x_0$ ,  $x_0 + \lambda(x - x_0) \in C$  for all  $\lambda > 0$ . Since the vertex  $x_0$  of the cone  $C$  will normally be obvious, we shall omit mentioning it.

†  $\text{co}\{z, z + \epsilon \delta z^1, \dots, z + \epsilon \delta z^k\}$  is the convex hull of  $z, z + \epsilon \delta z^1, \dots, z + \epsilon \delta z^k$ , i.e., the set of all points,  $y$ , of the form  $y = \mu_0 z + \mu_1(z + \epsilon \delta z^1) + \dots + \mu_k(z + \epsilon \delta z^k)$ , where  $\sum_{i=0}^k \mu_i = 1, \mu_i \geq 0$  for all  $i$ .

Whenever the radial cone  $RC(\hat{z}, \Omega)$  is a linearization of the first kind, it contains all the other linearization of the first kind of the set  $\Omega$  at  $\hat{z}$ . Consequently, in the various theorems to follow, the radial cone  $RC(\hat{z}, \Omega)$  should always be used if possible, since this will result in stronger necessary conditions.

Next, we define the  $C^{(1)}$  map  $F: E^n \rightarrow E^{m+1}$

$$F(z) = (f(z), r(z)).$$

We shall number the components of  $E^{m+1}$  from 0 to  $m$ , i.e.,  $y \in E^{m+1}$  is given by  $y = (y^0, y^1, \dots, y^m)$ . The Jacobian matrix of the map  $F(z)$ ,  $\left( \frac{\partial F^i(z)}{\partial z_j} \right)$ , will be denoted by  $\frac{\partial F(z)}{\partial z}$ .

For the Basic Problem stated above, the following theorem gives a necessary condition for optimality.

Theorem 1. If  $\hat{z}$  is an optimal solution to the Basic Problem, and  $C(\hat{z}, \Omega)$  is a linearization of the first kind of  $\Omega$  at  $\hat{z}$ , then there exists a nonzero vector  $\psi = (\psi^0, \psi^1, \dots, \psi^m) \in E^{m+1}$ , with  $\psi^0 \leq 0$ , such that for all  $\delta z \in \overline{C(\hat{z}, \Omega)}$  (the closure of  $C(\hat{z}, \Omega)$  in  $E^n$ )

$$(1) \quad \left\langle \psi, \frac{\partial F(\hat{z})}{\partial z} \delta z \right\rangle \leq 0.$$

Proof. Let  $K(\hat{z}) \subset E^{m+1}$  be the cone defined by

$$(2) \quad K(\hat{z}) = \frac{\partial F(\hat{z})}{\partial z} C(\hat{z}, \Omega)$$

$K(\hat{z})$  is convex because  $C(\hat{z}, \Omega)$  is convex and  $\frac{\partial F(\hat{z})}{\partial z}$  is a linear map. Let  $\hat{y} = F(\hat{z})$ . We shall now show that the cone  $\{\hat{y}\} + K(\hat{z})$  must be separated from the ray

$$(3) \quad R = \{y: y = \hat{y} + \beta(-1, 0, \dots, 0), \beta \geq 0\},$$

i.e. that there must exist a nonzero vector  $\psi \in E^{m+1}$  such that

$$(4) \quad (i) \quad \langle \psi, y - \hat{y} \rangle \leq 0 \text{ for every } y \in \{\hat{y}\} + K(\hat{z})$$

$$(ii) \quad \langle \psi, y - \hat{y} \rangle \geq 0 \text{ for every } y \in R.$$

Suppose that the cone  $\{\hat{y}\} + K(\hat{z})$  and the ray  $R$  are not separated. Then the cone  $K(\hat{z})$  must be of dimension  $m+1$  and  $R$  must be an interior ray of  $\{\hat{y}\} + K(\hat{z})$  (i.e., all points of  $R$  except  $\hat{y}$  are interior points of  $\{\hat{y}\} + K(\hat{z})$ ).

Let us now construct in the cone  $\{\hat{y}\} + K(\hat{z})$  a simplex  $\Sigma$  with vertices  $\hat{y}, \hat{y} + \delta y^1, \hat{y} + \delta y^2, \dots, \hat{y} + \delta y^{m+1}$  such that

- (i) there exists a point  $y$  on  $R$  (which we shall write as  $y = \hat{y} + \delta y^0, \delta y^0 = \gamma(-1, 0, \dots, 0)$  with  $\gamma > 0$ ), different from  $\hat{y}$ , which lies in the interior of  $\Sigma$ ,
- (ii) there exists a set of vectors  $\delta z^i \in C(\hat{z}, \Omega)$  satisfying

$$(5) \quad \delta y^i = \frac{\partial F(\hat{z})}{\partial z^i} \delta z^i, \quad i = 1, \dots, m+1$$

and such that

$$(6) \quad \text{co}\{\hat{z}, \hat{z} + \delta z^1, \dots, \hat{z} + \delta z^{m+1}\} \subset \Omega.$$

It is possible to satisfy (i) because  $R$  is an interior ray of the  $m+1$  dimensional cone  $\{y\} + K(z)$ , and it is possible to satisfy (ii) because  $C(\hat{z}, \Omega)$  is a linearization of the first kind. Note that the vectors  $\delta z^i$ ,  $i = 1, \dots, m+1$  are linearly independent since the vectors  $\delta y^1, \delta y^2, \dots, \delta y^{m+1}$  are linearly independent.

For  $0 < \alpha \leq 1$ , let  $S_\alpha \subset \Sigma$  be a sphere with center  $\hat{y} + \alpha \delta y^0$  and radius  $\alpha r$ , where  $r > 0$ . This sphere can always be constructed because  $\hat{y} + \delta y^0$  is an interior point of  $\Sigma$ . For a fixed  $\alpha$ ,  $0 < \alpha \leq 1$ , we now construct a map  $G_\alpha$  from  $S_\alpha - \{\hat{y} + \alpha \delta y^0\}$  into  $E^{m+1}$  as follows. For any  $x \in S_\alpha - \{\hat{y} + \alpha \delta y^0\}$ , let

$$(7) \quad G_\alpha(x) = F(\hat{z} + ZY^{-1}(\alpha \delta y^0 + x)) - (\hat{y} + \alpha \delta y^0)$$

where  $Y$  is a  $(m+1) \times (m+1)$  matrix whose  $i$ th column is  $\delta y^i$ ,  $i = 1, \dots, m+1$ , and  $Z$  is a  $n \times (m+1)$  matrix whose  $i$ th column is  $\delta z^i$ . The matrix  $Y$  is invertible because the  $\delta y^i$  form a basis for  $E^{m+1}$ , by construction.

Expanding the right hand side of (7) about  $\hat{z}$ , we get

$$G_{\alpha}(x) = \hat{y} + \frac{\partial F(\hat{z})}{\partial z} ZY^{-1} (\alpha \delta y^0 + x) - (\hat{y} + \alpha \delta y^0) + o(ZY^{-1}(\alpha \delta y^0 + x))$$

(8)

where  $o(\cdot)$  is a continuous function such that  $\lim_{\|y\| \rightarrow 0} \frac{\|o(y)\|}{\|y\|} = 0$ .  
By definition,  $\frac{\partial F(\hat{z})}{\partial z} Z = Y$ , and hence (8) simplifies to

$$G_{\alpha}(x) = x + o(ZY^{-1}(\alpha \delta y^0 + x))$$

(9)

Now, for  $x \in \partial(S_{\alpha} - \{\hat{y} + \alpha \delta y^0\})$  (the boundary of the sphere)  $\|x\| = \alpha r$  and we may write  $x = \alpha \rho_1$ , where  $\|\rho_1\| = r$ . Hence, for  $x \in \partial(S_{\alpha} - \{\hat{y} + \alpha \delta y^0\})$

$$G_{\alpha}(\alpha \rho_1) = \alpha \rho_1 + o(\alpha ZY^{-1}(\delta y^0 + \rho_1))$$

(10)

Consequently, there exists an  $\alpha^*$ ,  $0 < \alpha^* \leq 1$ , such that for all  $\rho_1 \in E^{m+1}$ , with  $\|\rho_1\| = r$ ,

$$\|o(\alpha^* ZY^{-1}(\delta y^0 + \rho_1))\| < \alpha^* r$$

(11)

We now conclude from Brouwer's Fixed Point Theorem that there exists a  $\tilde{x} \in S_{\alpha^*} - \{\hat{y} + \alpha^* \delta y^0\}$  such that

$$G_{\alpha^*}(\tilde{x}) = 0.$$

(12)

(i.e.  $\alpha \rho_1 = -o(\alpha(ZY^{-1}(\delta y^0 + \rho_1)))$  and hence  $\alpha \rho_1$  is a fixed point of  $-o(\alpha(ZY^{-1}(\delta y^0 + \rho_1)))$ .)

i. e.,

$$(13) \quad F(\hat{z} + ZY^{-1}(\alpha^* \delta y^0 + \tilde{x})) = \hat{y} + \alpha^* \delta y^0$$

Now  $\hat{y} + \alpha^* \delta y^0 = \text{col}(f(\hat{z}) - \alpha^* \gamma, 0, 0, \dots, 0)$ , where  $\gamma > 0$ . Thus, expanding (13),

$$(14) \quad r(\hat{z} + ZY^{-1}(\alpha^* \delta y^0 + \tilde{x})) = 0$$

and

$$(15) \quad f(\hat{z} + ZY^{-1}(\alpha^* \delta y^0 + \tilde{x})) = f(\hat{z}) = \alpha^* - \gamma < f(\hat{z})$$

Furthermore, because of (6) and the fact that for any  $\delta y$  in the simplex  $Z = \{\hat{y}\}$ , the vector  $x = \hat{z} + ZY^{-1} \delta y$  belongs to  $\text{co}\{\hat{z}, \hat{z} + \delta z^1, \dots, \hat{z} + \delta z^{m+1}\}$ ,

$$(16) \quad \hat{z} + ZY^{-1}(\alpha^* \delta y^0 + \tilde{x}) \in \Omega$$

Hence  $\hat{z}$  is not optimal, which is a contradiction. We therefore conclude that the cone  $\{\hat{y}\} + K(\hat{z})$  and the ray  $R$  must be separated, i. e., there must exist a nonzero vector  $\psi \in \mathbb{R}^{m+1}$  such that

$$(17) \quad (1) \quad \langle \psi, (y - \hat{y}) \rangle \leq 0 \quad \text{for every } y \in \{\hat{y}\} + K(\hat{z})$$

and

$$(18) \quad (ii) \quad \langle \psi, (y - \hat{y}) \rangle \geq 0 \quad \text{for every } y \in R$$

Substituting (2) in (17), we have

$$(19) \quad \left\langle \psi, \frac{\partial F(\hat{z})}{\partial z} \delta z \right\rangle \leq 0 \quad \text{for every } \delta z \in C(\hat{z}, \Omega)$$

Clearly, (19) must also hold for every  $\delta z \in \overline{C(\hat{z}, \Omega)}$ .

Substituting for  $y$  from (3) in (18), we have

$$(20) \quad \langle \psi, (-1, 0, \dots, 0) \rangle = -\psi^0 \geq 0.$$

This completes the proof.

It has been pointed out by Neustadt [2] that Theorem 4 remains valid under the relaxed assumption that  $C(\hat{z}, \Omega)$  is a linearization of the second kind of  $\Omega$  at  $\hat{z}$ , defined as follows.

**Definition.** A convex cone  $C(z, \Omega) \subset E^n$  will be called a linearisation of the second kind of the constraint set  $\Omega$  at  $z$ , if, for any finite collection  $\{\delta z^1, \delta z^2, \dots, \delta z^k\}$  of linearly independent vectors in  $C(z, \Omega)$ , there exists an  $\epsilon > 0$ , possibly depending on  $z, \delta z^1, \dots, \delta z^k$ , and a continuous map  $\zeta$  from  $\text{co}\{z, z + \epsilon \delta z^1, \dots, z + \epsilon \delta z^k\}$  into  $\Omega$ , such that  $\zeta(z + \delta z) = z + \delta z + o(\delta z)$ , where  $\lim_{\|\delta z\| \rightarrow 0} \frac{\|o(\delta z)\|}{\|\delta z\|} = 0$ .

Remark. We observe that if  $C(z, \Omega)$  is a linearization of the first kind of  $\Omega$  at  $z$ , then it is also a linearization of the second kind of  $\Omega$  at  $z$ , with the map  $\xi$  being the identity. Thus, unless we have specific cause to indicate whether a cone  $C(z, \Omega)$  is a linearization of the first or second kind, we shall refer to it simply as a linearization of  $\Omega$  at  $z$ . We now restate Theorem 1 in this form.

FUNDAMENTAL THEOREM

If  $\hat{z}$  is an optimal solution to the basic problem and  $C(\hat{z}, \Omega)$  is a linearization of  $\Omega$  at  $\hat{z}$ , then there exists a nonzero vector  $\psi = (\psi^0, \psi^1, \dots, \psi^m) \in E^{m+1}$  with  $\psi^0 \leq 0$ , such that for all  $\delta z \in \overline{C(\hat{z}, \Omega)}$ , (the closure of  $C(\hat{z}, \Omega)$  in  $E^n$ ),  $\langle \psi, \frac{\partial F(\hat{z})}{\partial z} \delta z \rangle \leq 0$ .

The reader may easily modify the proof of Theorem 1 so as to apply to Theorem 1'. Finally, it should be pointed out that all conditions such as continuity differentiability, etc., imposed on the various functions need only hold in a neighborhood of the optimal point.

We shall now show how a number of classical optimization problems can be cast in the form of the Basic Problem, and we shall then apply Theorem 1 or Theorem 1' to rederive several classical conditions for optimality, as well as to obtain some new ones.

### Classical Theory of Lagrange Multipliers

The classical constrained minimization problem admits equality constraints only. Thus, it is the Basic Problem with  $\Omega = E^n$ , the entire space. Clearly,  $E^n$  is a linearization of the first kind for  $E^n$  at any point  $z \in E^n$ .

Thus, we conclude from Theorem 1 that if  $z$  is an optimal solution of the Basic Problem, with  $\Omega = E^n$ , then there exists a nonzero vector  $\psi \in E^{m+1}$  such that

$$(21) \quad \left\langle \psi, \frac{\partial F(\hat{z})}{\partial z} \delta z \right\rangle \leq 0 \text{ for all } \delta z \in E^n$$

This may be rewritten as

$$(22) \quad \left\langle \frac{\partial F(\hat{z})}{\partial z}^T \psi, \delta z \right\rangle \leq 0 \text{ for all } \delta z \in E^n$$

Since for any  $\delta z \in E^n$ ,  $-\delta z$  is also in  $E^n$ , we conclude from (22) that

$$(23) \quad \frac{\partial F(\hat{z})}{\partial z}^T \psi = 0$$

Now,  $\frac{\partial F(\hat{z})}{\partial z}^T$  is a  $n \times (m+1)$  matrix with columns  $\nabla f(\hat{z})$ ,  $\nabla r^1(\hat{z})$ ,  $\dots$ ,  $\nabla r^m(\hat{z})$ , where  $\nabla f(\hat{z}) = \left( \frac{\partial f(\hat{z})}{\partial z_1}, \dots, \frac{\partial f(\hat{z})}{\partial z_n} \right)$ ,  $\nabla r^1(\hat{z}) = \left( \frac{\partial r^1(\hat{z})}{\partial z_1}, \dots, \frac{\partial r^1(\hat{z})}{\partial z_n} \right)$ . We may therefore expand (23) into the form

$$(24) \quad \psi^0 \nabla f(z) + \sum_{i=1}^m \psi^i \nabla r^i(z) = 0$$

We have thus reproved the following classical result.

Theorem 2. Let  $f(\cdot)$ ,  $r^1(\cdot)$ ,  $r^2(\cdot)$ ,  $\dots$ ,  $r^m(\cdot)$  be real valued, continuously differentiable functions on  $E^n$ . If  $\hat{z} \in E^n$  minimizes  $f(z)$  subject to the constraints  $r^i(z) = 0$ ,  $i = 1, 2, \dots, m$ , then there exist scalar multipliers,  $\psi^0, \psi^1, \dots, \psi^m$ , not all zero, such that the function  $H$  on  $E^n$  which they define by

$$(25) \quad H(z) = \psi^0 f(z) + \sum_{i=1}^m \psi^i r^i(z)$$

has a stationary point at  $z = \hat{z}$ , i.e., (24) is satisfied.

It is usual to assume that the gradient vectors  $\nabla r^i(z)$ ,  $i = 1, 2, \dots, m$ , are linearly independent for all  $z$  such that  $r(z) = 0$ . This precludes  $\sum_{i=1}^m \psi^i \nabla r^i(\hat{z}) = 0$  and hence in (24)  $\psi^0 \neq 0$ . Multiplying (24) by  $1/\psi^0$  and letting  $\hat{\lambda}_i = \psi^i/\psi^0$ ,  $i = 1, 2, \dots, m$ , we now deduce the more commonly seen condition.

Theorem 2'. If  $\hat{z}$  minimizes  $f(z)$  subject to  $r(z) = 0$ , and the gradients  $\nabla r_i(\hat{z})$ ,  $i = 1, 2, \dots, m$ , are linearly independent, then there exists a vector  $\lambda \in E^m$  such that the Lagrangian  $L$  on  $E^n \times E^m$ , defined by

$$(26) \quad L(z, \lambda) = f(z) + \sum_{i=1}^m \lambda^i r^i(z)$$

has a stationary point at  $(\hat{z}, \hat{\lambda})$ .

We note that by (24)  $\frac{\partial L(\hat{z}, \hat{\lambda})}{\partial z} = 0$  and that  $\frac{\partial L(\hat{z}, \hat{\lambda})}{\partial \lambda} = r(\hat{z}) = 0$ , by assumption.

### Nonlinear Programming

Let  $f: E^n \rightarrow E^1$ ,  $r: E^n \rightarrow E^m$ , and  $q: E^n \rightarrow E^k$  be continuously differentiable functions. The standard Nonlinear Programming Problem is that of minimizing  $f(z)$  subject to the constraints that  $r(z) = 0$  and  $q(z) \leq 0$ .

This corresponds to the special case of the Basic Problem, with  $\Omega = \{z: q(z) \leq 0\}$ . We shall now show how Theorem 1 can be used to obtain various commonly known necessary conditions for  $z$  to be optimal.

Given a particular point  $z \in \Omega$ , we shall often have occasion to divide the components of the inequality constraints functions,  $q^i$ ,  $i = 1, \dots, k$ , into two sets; those for which  $q^i(z) = 0$  and those for which  $q^i(z) < 0$ . To simplify notation we introduce the following definition.

Definition. For  $z \in \Omega$ , let the index set  $I(z)$  be defined by

$$(27) \quad I(z) = \{i : q^i(z) = 0\}$$

The constraints  $q^i$ ,  $i \in I(z)$  will be called the active constraints at  $z$ . We shall denote by  $\bar{I}(z)$  the complement of  $I(z)$  in  $\{1, \dots, k\}$ .

The set  $\Omega = \{z : q(z) \leq 0\}$  introduced above is assumed to satisfy the following condition:

Assumption (A1).<sup>†</sup> Let  $\hat{z} \in \Omega$  be an optimal solution<sup>of</sup> the Nonlinear Programming Problem. Then, there exists a vector  $h \in E^n$  such that

$$\langle \nabla q^i(\hat{z}), h \rangle < 0 \text{ for all } i \in I(\hat{z})$$

A sufficient condition for (A1) to be satisfied is that the vectors  $\nabla q^i(\hat{z})$ ,  $i \in I(\hat{z})$  be linearly independent (see Corollary to Lemma 3).

Definition: For any  $z \in \Omega$ , the internal cone of  $\Omega$  at  $z$ , denoted by  $IC(z, \Omega)$ , is defined by

$$IC(z, \Omega) = \{\delta z : \langle \nabla q^i(z), \delta z \rangle < 0 \text{ for all } i \in I(z)\}$$

<sup>†</sup> When some of the functions  $q^i$ ,  $i \in I(z)$ , are linear, it suffices to require that there exist a vector  $h \in E^n$  such that  $\langle \nabla q^i(z), h \rangle \leq 0$  for these functions and  $\langle \nabla q^i(z), h \rangle < 0$  for the remaining functions  $q^i$ ,  $i \in I(z)$ .

By assumption (A1), the convex cone  $IC(z, \Omega)$  is nonempty. It is a simple exercise in the use of Taylor's Theorem to prove the following lemma.

Lemma 1. If  $IC(z, \Omega) \neq \emptyset$ , the empty set, then

(i)  $IC(z, \Omega)$  is a linearization of the first kind of  $\Omega$  at  $z$ ,

(ii)  $\overline{IC(z, \Omega)} = \{\delta z : \langle \nabla q^i(z), \delta z \rangle \leq 0 \text{ for all } i \in I(z)\}$

When specialized to the Nonlinear Programming Problem, Theorem 1' assumes the following form.

Theorem 3. If  $z$  is an optimal solution to the Nonlinear Programming Problem, with (A1) satisfied, then there exists a nonzero vector  $\psi \in E^{m+1}$ , with  $\psi^0 \leq 0$ , such that for all  $\delta z \in \overline{IC(z, \Omega)} = \{\delta z : \langle \nabla q^i(z), \delta z \rangle \leq 0 \text{ for all } i \in I(z)\}$ ,

$$\left\langle \frac{\partial H(z)}{\partial z}, \delta z \right\rangle \leq 0$$

where  $H(z) = \psi^0 f(z) + \sum_{i=1}^m \psi^i r^i(z)$ .

Using Theorem 3 and Farkas Lemma we obtain the following necessary condition for optimality, which is in a form more familiar to specialists in mathematical programming.

Theorem 4. If  $\hat{z}$  is an optimal solution to the Nonlinear Programming Problem, with (A1) satisfied, then there exist a nonzero vector  $\psi \in E^{m+1}$ , with  $\psi^0 \leq 0$ , and a vector  $\mu \in E^k$ , with  $\mu \leq 0$ , such that

$$(i) \quad \psi^0 \nabla f(\hat{z}) + \sum_{i=1}^m \psi^i \nabla r^i(\hat{z}) + \sum_{i=1}^k \mu^i \nabla q^i(\hat{z}) = 0$$

and

$$(ii) \quad \sum_{i=1}^k \mu^i q^i(\hat{z}) = 0$$

Proof. From Theorem 3,

$$\left\langle \frac{\partial H(\hat{z})}{\partial z}, \delta z \right\rangle \leq 0$$

for all  $\delta z$  such that  $\langle \nabla q^i(\hat{z}), \delta z \rangle \leq 0$ ,  $i \in I(\hat{z})$ .

By Farkas Lemma, there exist scalars  $\mu^i \leq 0$ ,  $i \in I(\hat{z})$  such that

$$\frac{\partial H(\hat{z})}{\partial z} + \sum_{i \in I(\hat{z})} \mu^i \nabla q^i(\hat{z}) = 0$$

Let  $\mu^i = 0$  for  $i \in \overline{I(\hat{z})}$ . This completes the proof.

Most of the other well-known necessary conditions for Nonlinear Programming Problems can be obtained from Theorem 4 by making additional assumptions on the functions  $r$  and  $q$ . For example, the following corollaries to Theorem 4 are immediate consequences of that theorem.

Corollary 1. If assumption (A1) is satisfied and the vectors  $\nabla r^i(\hat{z})$ ,  $i = 1, \dots, m$ , are linearly independent, then there exist vectors  $\psi \in E^{m+1}$ ,  $\mu \in E^k$  which satisfy the conditions of Theorem 4 and such that  $(\psi^0, \mu) \neq 0$ .

Corollary 2. If  $\nabla r^i(\hat{z})$ ,  $i = 1, \dots, m$ , together with  $\nabla q^i(\hat{z})$ ,  $i \in I(\hat{z})$ , are linearly independent vectors, there exists a vector  $\psi \in E^{m+1}$  satisfying the conditions of Theorem 4 with  $\psi^0 < 0$ .

The assumption in Corollary 2 is a well-known [11] sufficient condition for the Kuhn-Tucker constraint qualification to be satisfied. When it is added to Theorem 4 we obtain a slightly restricted form<sup>†</sup> of the Kuhn-Tucker Theorem [4].

Corollary 3. If there exists a vector  $h \in E^n$  such that  $\langle \nabla q^i(\hat{z}), h \rangle < 0$  for all  $i \in I(\hat{z})$ ,  $\langle \nabla r^i(\hat{z}), h \rangle = 0$  for  $i = 1, \dots, m$ , and the vectors  $\nabla r^i(\hat{z})$ ,  $i = 1, \dots, m$ , are linearly independent, then there exists a vector  $\psi \in E^{m+1}$  satisfying the conditions of Theorem 4 with  $\psi^0 < 0$ .

The assumption in this corollary is a sufficient condition for the weakened constraint qualification [13] to be satisfied. Augmented by this assumption, Theorem 4 becomes a slightly restricted form<sup>†</sup> of the Kuhn-Tucker Theorem with the weakened constraint qualification.

<sup>†</sup> In practice, the Kuhn-Tucker constraint conditions can rarely be shown to be satisfied unless the restrictions imposed in Corollaries 2 and 3 hold.

### III THE MAXIMUM PRINCIPLE

To illustrate the applicability of the theory just developed, we shall use it to obtain the Pontryagin Maximum Principle for optimal control problems.

Consider a dynamical system described by the differential equation:

1. 
$$\frac{dx}{dt} = \underline{h}(x,u)$$

for all  $t$  in the compact interval  $I = [t_1, t_2]$ , where  $x(t) \in E^n$  is the state of the system at time  $t$ ,  $u(t) \in E^m$  is the input or control of the system at time  $t$ , and  $\underline{h}$  is a function defined on  $E^n \times E^m$  with range in  $E^n$ .

The Fixed Time Optimal Control Problem is that of finding a control  $u(t)$ ,  $t \in I$ , and a corresponding trajectory  $x(t)$ , determined by ( ), such that

2. for  $t \in I$ ,  $u(t)$  is a measurable, essentially bounded function

whose range is contained in an arbitrary but fixed subset  $U$  of  $E^m$ ;

3.  $\hat{x}(t_1) = x_0$ , where  $x_0$ , a fixed vector in  $E^n$ , is the given initial condition;

4.  $\hat{x}(t_2) \in X_2$ , where  $X_2 = \{x \in E^n | g(x) = 0\}$ , and  $g$  maps  $E^n$  into  $E^l$  ( $X_2$  is the fixed target set);

5. for every control  $u(t)$ , and corresponding trajectory  $x(t)$ , satisfying the conditions (2), (3), and (4),

$$\int_{t_1}^{t_2} f^0(x(t), u(t)) dt \geq \int_{t_1}^{t_2} f^0(\hat{x}(t), \hat{u}(t)) dt$$

where  $f^0$  (2.) is a cost function mapping  $E^n \times E^m$  into  $E^p$ .

We make the following assumptions:

6. The functions  $h(\cdot, \cdot)$  and  $f^0(\cdot, \cdot)$  are continuous in both  $x$  and  $u$ , and are continuously differentiable in  $x$ ;

7. The function  $g(\cdot)$  is continuously differentiable and the corresponding Jacobian matrix  $\frac{\partial g(x)}{\partial x}$  is of maximum rank for every  $x$  in  $X_2$ .

To transcribe the control problem into the form of the Basic Problem we require the following definitions:

Let  $I_\alpha$  denote the  $\alpha \times \alpha$  identity matrix and let  $0_{\alpha\beta}$  denote the  $\alpha \times \beta$  zero matrix. We define the projection matrices  $P_1$  and  $P_2$  as

$$8. P_1 = (I_n, 0_{n,n}) = (1, 0, 0, \dots, 0)$$

and

$$9. P_2 = (0_{n,n}, I_n) = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 0 \end{pmatrix}$$

Let  $h: E^{n+m} \times E^m \rightarrow E^{n+m}$  be the function defined by

$$10. h(z, u) = (f(P_1 z, u), h(P_2 z, u)), \quad z \in E^{n+m}, \quad u \in E^m.$$

Now consider the differential equation

$$11. \frac{dz}{dt} = h(z, u)$$

for some  $u(t) \in E^m$  for  $t \in I$ .

It is clear that the optimal control problem is equivalent to the problem of finding a control  $\hat{u}(t)$ ,  $t \in I$  and a corresponding trajectory  $z(t)$ , determined by (11), such that

(i) for  $t \in I$ ,  $\hat{u}(t)$  is a measurable, essentially bounded function, whose range is contained in an arbitrary but fixed subset  $U$  of  $E^m$ ;

(ii)  $\hat{z}(t_1) = (0, x_0) = z_0$ ; where  $x_0$ , a fixed vector in  $E^n$ , is the given initial condition;

14.  $\hat{z}(t_2) \in X_2^1$ , where  $X_2^1 = \{z \in E^{n+1} \mid g(P_2 z) = 0\}$ , where  $g$  maps  $E^n$  into  $E^l$ ;

15. for every control  $u(t)$ , with  $t \in I$ , and corresponding trajectory  $z(t)$ , satisfying (11) and the conditions (12), (13), and (14) above,

$$P_1 z(t_2) \geq P_1 \hat{z}(t_2)$$

Finally, we complete the transcription of the optimal control problem into the form of the Basic Problem by defining

16.  $f(z) = P_1 z(t_2)$ ,

17.  $r(z) = g(P_2 z(t_2))$ ,

by letting  
18. and we let  $\Omega$  be the set of all absolutely continuous functions  $z$  from  $I$  into  $E^{n+1}$  which, for some measurable, essentially bounded function  $u$  from  $I$  into  $U \subset E^m$ , satisfy the differential equation (11) for almost all  $t$  in  $I$ , with  $z(t_1) = (0, x_0)$ .

19. Remark: It is clear that with  $f$ ,  $r$ , and  $\Omega$  defined as in (16), (17), and (18), respectively, we have transcribed the optimal control problem into the form of the Basic Problem ~~1.1.1~~<sub>A</sub>. We shall call the transcribed optimal control "the optimal control problem in standard form."

We still have not defined the linear topological vector space  $\mathcal{L}$ . From (18) it is clear that  $\Omega$  is a subset of the linear space of all absolutely continuous functions from  $I$  into  $E^{n+1}$ . However, since we wish to use a linearization constructed first by Pontryagin et al. [ ], we find it necessary to imbed  $\Omega$  into a larger linear topological

space which we define below.

Let  $\mathcal{U}$  be the set of all upper semi-continuous real valued functions\* defined on  $I$ , and let  $\mathcal{J} = \mathcal{U} - \mathcal{U}$ . From the properties of upper and lower semi-continuous functions it follows that  $\mathcal{J}$  is a linear vector space. We then define  $\mathcal{L}$  to be the Cartesian product  $\mathcal{J}^{n+1} = \mathcal{J} \times \mathcal{J} \times \cdots \times \mathcal{J}$ , with the pointwise topology, [15] i.e., the topology which is constructed from the sub-base consisting of the family of all subsets of the form  $\{f \in \mathcal{L} : f(t) \in N\}$ , where  $t$  is a point in  $I$  and  $N$  is an open set in  $E^{n+1}$ .

It is easy to show that  $\hat{f}$  and  $r$ , as respectively defined by (16) and (17), are continuous on  $\mathcal{L}$ .

Let  $\hat{z}(t)$ , corresponding to the control  $\hat{u}(t)$ , be an optimal solution to the optimal control problem in standard form (19). We now proceed to construct a linearization for the constraint set  $\Omega$  at  $\hat{z}$ .

Let  $I_1 \subset I$  be the set of all points  $t$  at which  $\hat{u}(t)$  is regular, i.e.,

$$20. I_1 = \{t \mid t_1 < t < t_2, \lim_{\text{meas}(T) \rightarrow 0} \frac{\text{meas}(\hat{u}^{-1}(N) \cap T)}{\text{meas}(T)} = 1, \text{ for every}$$

neighborhood  $N$  of  $\hat{u}(t)$ ,  $t \in T \subset I\}$ .

\* Definition: A real valued function  $f: E^1 \rightarrow E^1$  is called upper semi-continuous at a point  $t_0$  in  $E^1$ , if  $\limsup_{t \rightarrow t_0} f(t) \leq f(t_0)$ . And it is called lower semi-continuous if  $-f$  is upper semi-continuous [ ].

Let  $\Phi(t, \tau)$  be the  $(n+1) \times (n+1)$  matrix which satisfies the linear differential equation

$$21. \frac{d}{dt} \Phi(t, \tau) = \frac{\partial h}{\partial z}(\hat{z}(t), \hat{u}(t)) \Phi(t, \tau)$$

for almost all  $t \in I$ , with  $\Phi(\tau, \tau) = I_{n+1}$ , the  $(p+n)$  identity matrix.

For any  $s \in I_1$  and  $v \in U$  we define

$$22. \delta z_{s,v}(t) = \begin{cases} 0 & \text{for } t_1 \leq t < s \\ \Phi(t, s)[h(\hat{z}(s), v) - h(\hat{z}(s), \hat{u}(s))], & s \leq t \leq t_2. \end{cases}$$

and

$$23. C(\hat{z}, \Omega) = \left\{ \delta z \in \mathcal{L} \mid \delta z(t) = \sum_{i=1}^k \alpha_i \delta z_{s_i, v_i}(t), \{s_1, s_2, \dots, s_k\} \subset I_1, \right.$$

$$\left. \{v_1, v_2, \dots, v_k\} \subset U, \alpha_i \geq 0, \text{ for } i=1, 2, \dots, k, k \text{ arbitrary finite} \right\}.$$

Because of its complexity, we shall delay the proof that the set  $C(\hat{z}, \Omega)$  defined in (23), is a linearization for the set  $\Omega$  at  $\hat{z}$  until the next section. The linear maps  $f'(\hat{z})$  and  $r'(\hat{z})$ , with  $F'(\hat{z}) = (f'(\hat{z}), r'(\hat{z}))$ , which we use with this linearization are

defined as follows. For every  $\delta z \in \mathcal{L}$ ,

$$f'(\hat{z})(\delta z) = P_1 \delta z(t_2)$$

and

$$25. \quad r'(\hat{z})(\delta z) = \frac{\partial g(P_2 \hat{z}(t_2))}{\partial x} P_2 \delta z(t_2).$$

Therefore, from Theorem (II.6) there exist a vector  $\mu$  in  $E^p$  and a vector  $\eta$  in  $E^l$  such that

$$\mu^i \leq 0 \text{ for } i=1, 2, \dots, p;$$

$$27. \quad (\mu, \eta) \neq 0;$$

$$28. \quad \langle \mu, P_1 \delta z(t_2) \rangle + \langle \eta, \frac{\partial g(P_2 \hat{z}(t_2))}{\partial x} P_2 \delta z(t_2) \rangle \leq 0 \text{ for all } \delta z \in \overline{C(\hat{z}, \Omega)}.$$

Since every  $\delta z_{s,v}(t)$  defined in (22), is in  $\overline{C(\hat{z}, \Omega)}$ , (28) implies that

$$29. \quad \langle \mu, P_1 \Phi(t_2, s) [h(\hat{z}(s), v) - h(\hat{z}(s), \hat{u}(s))] \rangle + \\ + \langle \eta, \frac{\partial g(P_2 \hat{z}(t_2))}{\partial x} P_2 \Phi(t_2, s) [h(\hat{z}(s), v) - h(\hat{z}(s), \hat{u}(s))] \rangle \leq 0$$

for every  $s \in I$ , and  $v \in U$ .

Hence,

$$30. \quad \langle \Phi^T(t_2, t) \left[ P_1^T \mu + P_2^T \frac{\partial g^T(P_2 \hat{z}(t_2))}{\partial x} \eta \right], h(\hat{z}(t), v) - h(\hat{z}(t), \hat{u}(t)) \rangle \leq 0$$

for every  $t \in I_1$ , and  $v \in U$ . Let  $4(\cdot) = (4^0(\cdot), 4^1(\cdot), \dots, 4^n(\cdot))$  into  $E^{n+1}$  defined by

$$31. \quad \psi(t) = \Phi^T(t_2, t)(P_1^T \mu + P_2^T \left( \frac{\partial g(P_2 \hat{z}(t_2))}{\partial x} \right)^T \eta),$$

i.e. for almost all  $t$  in  $I$ ,  $\psi(t)$  satisfies the differential equation:

$$32. \quad \frac{d}{dt} \psi(t) = - \left( \frac{\partial h(\hat{z}(t), \hat{u}(t))}{\partial x} \right)^T \psi(t) \quad ; \quad \psi(t_2) = P_1^T \mu + P_2^T \left( \frac{\partial g(P_2 \hat{z}(t_2))}{\partial x} \right)^T \eta.$$

Combining (30) and (31), we obtain

$$33. \quad \langle \psi(t), h(\hat{z}(t), \hat{u}(t)) \rangle = \text{Maximum} \{ \langle \psi(t), h(\hat{z}(t), v) \rangle \mid v \in U \} \text{ for } t \in I_1.$$

34. Since  $\text{meas}(I_1) = \text{meas}(I)$ , (33) holds for almost all  $t$  in  $I$ .

34. Remark: By assumption (see (79)),  $\frac{\partial g(P_2 \hat{z}(t_2))}{\partial x}$  is of maximum rank, and since  $(\mu, \eta) \neq 0$ ,  $\psi(t)$  as a solution to (31') is not identically zero.

Thus, we have proved the following theorem, which we state in

terms of the original quantities defining the optimal control problem,

and in which we shall substitute  $(p^0, p)$ , with  $p \in E^n$  as above.

## II. INFINITE DIMENSIONAL PROBLEMS

We shall now show how the Fundamental Theorem presented in Part I can be extended to problems in infinite dimensional spaces. As an application, we shall use our extension of the Fundamental Theorem to derive the Pontryagin Maximum Principle [2], for fixed time optimal control problems.

First let us formulate the equivalent of the Basic Problem in an infinite dimensional space.

1. BASIC PROBLEM: Let  $L$  be a linear topological space.

Given a function  $f(\cdot)$  mapping  $L$  into the reals, a function  $(\cdot)$  mapping  $L$  into  $E^m$  and a subset  $\Omega \subset L$ , find a vector  $\hat{X} \in \Omega$  such that  $e(\hat{X}) = 0$  and such that for all  $X \in \Omega$  satisfying  $r(X) = 0$ ,

$$2. \quad f(\hat{X}) \leq f(X)$$

We shall call any  $\hat{X}$  with the above properties an optimal solution.

The only difference between the formulations in the proceeding section and (1) above is that before we specified that the functions  $f$  and  $r$  are differentiable, which we do not do in (1), since differentiability is not a well defined concept in a general linear topological space.

However, to obtain an extension of the Fundamental Theorem we need a linear function from  $L$  into  $E^{m+1}$  to take the place of the Jacobian matrix  $\frac{\partial F(\bar{u})}{\partial \bar{u}}$  in (I.1) and a suitable continuous function from  $L$  into  $E^{m+1}$  to take the place of the function  $O(\cdot)$  in (I.9). Since we can no longer ensure the existence of such functions simply by requiring that  $f(\cdot)$  and  $r(\cdot)$  be differentiable, we take care of this situation by incorporating the required functions into the definition of a conical approximation. As we shall see later, this is not a restrictive practice.

3. DEFINITION: A convex cone  $C(\hat{x}, \Omega) \subset L$  will be called a conical approximation to the set  $\Omega$  at  $\hat{x} \in \Omega$ , with respect to the map  $F^A(f, r)$ , if there exists a linear function  $F'(\hat{x})(\cdot)$  from  $L$  into  $E^{m+1}$  such that for any finite collection  $\{\delta x_1, \delta x_2, \dots, \delta x_k\}$  of linearly independent vectors in  $C(\hat{x}, \Omega)$  there exists an  $\epsilon > 0$  a continuous map  $\delta(\cdot)$  from  $\text{co}\{\hat{x}, \hat{x} + \epsilon \delta x_1, \hat{x} + \epsilon \delta x_2, \dots, \hat{x} + \epsilon \delta x_k\}$  into  $\Omega$ , and a continuous map  $o(\cdot)$  from  $L$  into  $E^{m+1}$ , with  $\epsilon$ ,  $\delta$ , and  $o$  possibly depending on  $\hat{x}, \delta x_1, \delta x_2, \dots, \delta x_k$  which satisfy

$$4. \quad \lim_{\epsilon \rightarrow 0} \frac{\|o(y)\|}{\epsilon} = 0$$

uniformly, for all  $y \in \text{co}\{\hat{x}, \hat{x} + \epsilon \delta x_1, \dots, \hat{x} + \epsilon \delta x_k\}$ , and,

$$5. \quad F(\delta(x)) = F(\hat{x}) + F'(\hat{x})(x - \hat{x}) + o(x - \hat{x})$$

for all

$$x \in \text{co}\{\hat{x}, \hat{x} + \epsilon \delta x_1, \dots, \hat{x} + \epsilon \delta x_k\}.$$

We are now ready to extend theorem (2.3.9).

6. THEOREM: If  $\hat{x}$  is an optimal solution to the Barre problem  $(\cdot)$  and  $c(\hat{x}, \Omega)$  is a conical approximation to  $\Omega$  at  $\hat{x} \in \Omega$  with respect to the map  $F = (f, r)$ , then there exists a nonzero vector  $\psi = (\psi^0, \psi^1, \dots, \psi^m)$  in  $E^{m+1}$ , with  $\psi^0 \leq 0$ , such that  $\langle \psi, F'(\hat{x})(\delta x) \rangle \leq 0$  for all  $\delta x \in \overline{C(\hat{x}, \Omega)}$ , where  $\overline{C(\hat{x}, \Omega)}$  is the closure of  $C(\hat{x}, \Omega)$  in  $L$ .

### REFERENCES

1. M. Canon, C. Cullum, E. Polak, "Constrained Minimization Problems in Finite Dimensional Spaces". J SIAM Control, 4 (1966), 528-547.
2. L.W. Neustadt, "An Abstract Variational Theory with Applications to a Broad Class of Optimization Problems", (Part I, General Theory, J SIAM, 4 (1966) pp 505-527, Part II, Applications J SIAM 5 (1967) pp 90-137.
3. Pontryagin et al, "The Mathematical Theory of Optimal Processes". Interscience, 1962.

Lectures on  
**MATHEMATICAL PROGRAMMING**

by  
**GEORGE B. DANTZIG**

at the  
**American Mathematical Society Summer Seminar**  
on the  
**Mathematics of the Decision Sciences**  
**Stanford University**  
**July - August 1967**

LARGE-SCALE SYSTEM OPTIMIZATION: A REVIEW

by

George B. Dantzig  
Operations Research Center  
University of California, Berkeley

March 1965

ORC 65-9

This research has been partially supported by the Office of Naval Research under Contracts Nour-222(83) and Nour-3656(02) with the University of California. Reproduction in whole or part is permitted for any purpose of the United States Government.

## Large-Scale System Optimization: A Review

by

George B. Dantzig

Mathematical programming is a generic term for the related fields of Linear Programming, Network Flow Theory, Integer Programming, Convex and Non-Linear Programming, and Programming under Uncertainty. Its research has problems, particularly those problems where random events and decision events occur alternately in successive stages. In problems where such uncertainty occurs, what is usually done in formulating is to replace the uncertain elements with their expected values (with possible an added safety factor). It is well known that a plan based on expected values of its coefficients and constraints can lead to answers that are not correct. Although the use of expected value does not lead to the best answer, it is entirely possible that it could lead to excellent plans indistinguishable from the optimum in the run-of-the-mill application. When one considers instead, a direct attack on uncertainty via mathematical programming, it inevitably leads to the consideration of large-scale systems. These, because of their structure, have proven difficult of solution so far, but, I believe, of intensive investigation in the future.

Mathematical programming is a term invented by Robert Dorfman of Harvard around 1950. He felt that at that time, the fundamentals of linear programming were well enough known that the wave-of-the-future lay in the extension of the methods of linear programming into the non-linear programming field. Certainly we today, 15 years later, feel this is true. In the Calculus, the derivative (or first order approximation) plays a key role. Applied to non-linear inequality systems, it leads to approximation by linear inequality systems. This is one way which these extensions have taken place, and illustrates why the various fields

comprised under mathematical programming are related. Here are some other ways:

One attempts to extend the concept of duality to nonlinear systems. Having done so, one tries to combine the combinatorial power of linear programs with the classical steepest descent processes to solve non-linear programs.

One attempts, as we have just noted, to reduce problems involving uncertainty to equivalent deterministic systems and to large-scale systems with special structure.

One tries to solve an integer program by replacing it with an equivalent linear program; that is to say, by cleverly building up a set of linear inequalities that are both necessary and sufficient.

In all of these developments, one characteristic stands out; namely in one way or another, techniques for solving large-scale systems play a dominant role.

Accordingly, let us look first at direct methods for handling large problems. Around 1954 or so, under the auspices of The RAND Corporation, William Orchard-Hays produced the first truly commercial linear programming code. It had many features that helped amateurs to get their problem on the machine with a reasonable chance of getting an answer. Today, the building of a linear programming code (complete with all the special features) is a major undertaking which is expensive to produce and to maintain.

As applications grow, there has been an increasing demand to handle truly enormous systems. The Russian, Kantorovich, in his 1939 pamphlet, envisioned such a possibility. Already, linear programming models of industrial systems have been solved with more than  $10^6$  variables and  $10^4$  equations. These models, of course, do not have general matrix structure and it is not likely that any instance of a large practical problem will ever have general structure. The reason is obvious. Just imagine the physical task alone of finding all the coefficients for a thousand by ten thousand general linear programs (there could be as high as  $10^7$  non-

zero coefficients).

Fortunately, large-scale practical models tend to have a low percentage of non-zero coefficients; in fact, under 5%, sometimes under 1%. Orchard-Hays' first code exploited this characteristic by making use of a "pricing vector". This made inexpensive the selection of the pivot column directly from the original non-dense data. The pivot column here refers to one of the steps in the simplex method for solving linear programs.

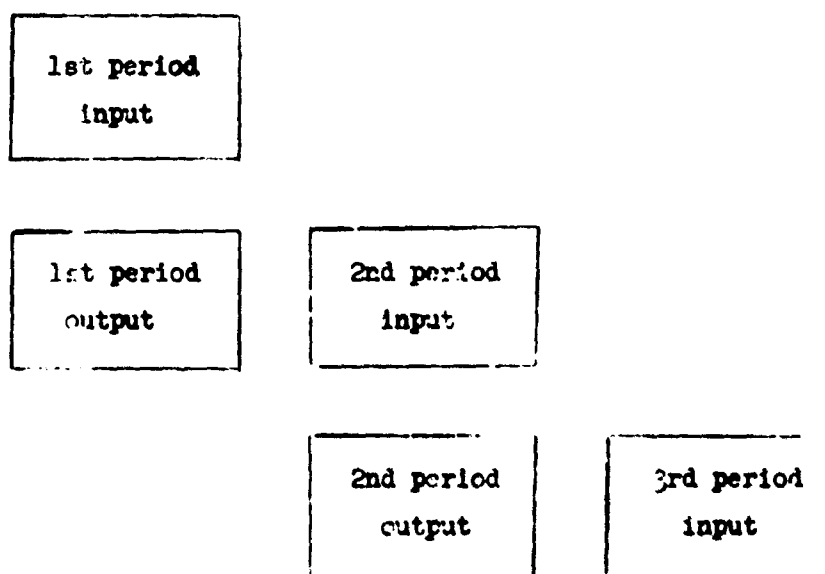
As systems have grown in size, every advantage has also been taken of the characteristics of the improved computers. It has been discovered recently that the size of the inverse representation of the basis in the simplex method could have an important effect on running time. Therefore, compact-inverse schemes along the lines first proposed by Harry Markowitz of RAND have become increasingly important. Recently, two groups working independently, developed this approach with astounding results. For example, the Standard Oil Company of California group reports running-time on some of their typical large problems cut to 1/4.

How to find the most compact inverse representation of a sparse matrix is still an unsolved problem:

Conjecture: If a non-singular matrix has  $K$  non-zero elements, it is always possible to represent them as a product of elementary matrices such that the total number of non-zero entries (excluding their diagonal unit elements) is at most  $K$ . [Incidentally, the empirical schemes just mentioned often have no more than  $K + 10\%K$  non-zeros in the inverse representation.]

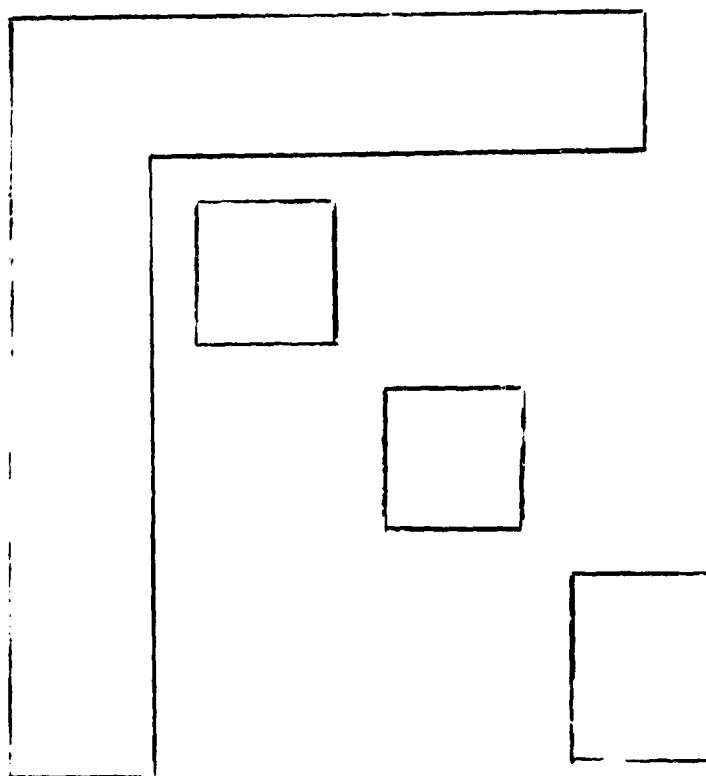
Dynamic structures are interesting in themselves, and could have important applications. One such is the linear control processes proposed by Pontryagin.

(I will speak of his problem later in connection with the decomposition principle.) As early as 1954, I published a paper on how to compact the inverse representation of the basis with a staircase structure (see figure). Again in 1962, I discussed another method which permitted one to find a compact inverse and then efficiently maintain this compactness in moving from one iteration to the next. There have been several other proposals, all excellent, that seek to apply the simplex method to the full system by compacting the inverse. As far as I know, none of these direct proposals have been realized in computer codes.



Large-scale systems have been attacked indirectly by means of the decomposition principle. Several codes have been written, and some of the recent experiences have been encouraging. J. F. Benders in his thesis "Partitioning in Mathematical Programming 1960", developed the dual of the decomposition principle, and shows how this approach can be used to deal with the mixed-integer programming problem. Rosen and Peale have each proposed partitioning methods

for solving systems whose structures fit into the framework of a common horizontal and/or vertical border while the remainder is a diagonal set of independent blocks.



Purely combinatorial problems form an important division of mathematical programming. They fall briefly into two categories. The first are those problems whose structures are special -- like the transportation problem -- or the minimum number of arcs which "cover" nodes in a network (graph). For these, special methods have been sought.

One of the most tantalizing problems of this type has been the travelling salesman problem. It is so close to a network-flow type problem that one would hope to find some easy representation of the faces of its polyhedral solution set. So far, none has been discovered. There is also a close relation between

covering problems and the famous four-color problem. The other approach to combinatorial problems is through integer programming. This was first used in 1954 to solve a particular large-scale travelling salesman problem by Fulkerson, Johnson, and myself. In 1958, Gomory laid the foundations of this field by showing how to systematically determine a necessary and sufficient system of linear inequalities.

The inter-relation between large-scale system methods and integer programming was brought out in a recent paper of Gomory entitled "Large and Non-convex Problems in Linear Programming". Here, Gomory reviews in a unified manner, how the ideas of integer programming and those of the decomposition principle can be combined to solve many important applications such as the paper-trim problem, multi-commodity flows in networks, programming of economic lot sizes, etc.

Integer programming methods are being experimented with in a number of places. It seems likely that we are nearing a threshold, and that we will soon see some excellent commercial codes produced and used successfully for certain problems.

I will not, in this presentation, describe the developments in non-linear programming. Rather, I have chosen to illustrate the power of certain non-linear programming ideas, such as the generalized linear program of Wolfe to an interesting problem in linear control theory.

But first, I would like to review the concept of a Generalized Program. This differs from an ordinary linear program. Instead of coefficients in each column being known, the column  $P_j$  may be freely drawn from a convex set,  $C_j$ .

PROBLEM: Find  $\text{Min } x_0$ ,  $\{x_j \geq 0, p_j \in C_j\}$ ,  $p_0 \in C_0, q \in C_q$  such that

$$p_0 x_0 + \dots + p_n x_n = q.$$

As an example, consider the CONVEX PROGRAM: Find  $(x_1, x_2, \dots, x_n) \in C$  compact and convex such that

$$\phi_1(x) \leq 0$$

$$\phi_2(x) \leq 0$$

.

.

.

$$\phi_n(x) \leq 0$$

$$\phi_0(x) = Z(\text{Min})$$

where  $\phi_i(x)$  are continuous convex functions of  $x \in C$ . Let us assume an  $x = x^0$  is known such that  $\phi_i(x^0) < 0$  for  $i \neq 0$ . It can be shown that the generalized program on the following page is equivalent.

PROBLEM: Find Min  $Z$  and  $\lambda, \lambda_i \geq 0, \mu_i \geq 0$  such that

				<u>Prices</u>	
1		1		= 1	$\pi_0$
$\phi_1(x)$	$\lambda +$	$\phi_1(x^0)$	$\lambda_0 + \mu_1$	= 0	$\pi_1$
.		.		= 0	$\pi_2$
.		.	$\mu_2$		
.		.	.		
.		.	.		
$\phi_m(x)$		$\phi_m(x^0)$	$\mu_m$	= 0	$\pi_m$
$\phi_0(x)$		$\phi_0(x^0)$		(-Z) = 0	1
<div style="display: flex; justify-content: space-around; align-items: center;"> <span>•</span> <span>•</span> <span>•</span> <span>..</span> <span>•</span> <span>•</span> </div>					

The position of an initial basic-set of columns is indicated by heavy dots.

The associated set of simplex multipliers are denoted by  $\pi_i$ ; initially,

$\pi = \pi^0$  where  $\pi_1^0 = 0$  for  $i \neq 0$  and  $\pi_0^0 = -\phi_0(x^0)$ . The next step is to form

the Lagrangian

$$\phi(X) = \phi_0(x) + \sum_{i=1}^m \pi_i \phi_i(x) + \pi_0$$

for  $\pi = \pi^0$  and to minimize  $\phi(x)$  for  $x \in C$ . Notice that the Lagrangian is

precisely what we get if we were to "price out" the general column of coeffi-

cients of the variable  $\lambda$  using the price vector to form the inner product.

Thus, we wish to choose  $x$  such that this scalar product is minimum.

Let  $x = x^0$  be the value of the minimizing  $x$ . If  $\phi(x^0) = 0$ , then

$x^0$  is an optimum solution. If  $\phi(x^0) < 0$ , an extra column is inserted in the

generalized program with coefficients  $[1, \phi_1(x^1), \dots, \phi_m(x^1), \phi_0(x^1)]$  and vari-

able  $\lambda_1$ . The problem, restricted to those variables whose columns have known coefficients, is then optimized using the simplex method. This gives rise to a new set of simplex multipliers  $\pi = \pi'$ . This gives rise to a new solution  $x = x^2$ , etc. At any stage, the approximate solution is  $x = \sum \lambda_i x^i$  using for  $\lambda_1$  those  $\lambda_i$  which solve the problem restricted to those variables with known coefficients.

Let us turn to a problem in control theory. The application of mathematical programming methods to solve control problems has been studied by Zadeh and Whalen, by Ben rosen, and others. I would like to confine myself, however, to Linear Control Theory as described by Pontryagin, Botlyanski, Gamkrelidze, and Mischenko in Chapter III of their book on this subject.

We consider an "object" defined by its  $n + 1$  coordinates  $x = (\xi_0, \xi_1, \dots, \xi_n)$  whose "motion", described as a function of a "time" parameter  $t$ , can be written as a linear system of differential equations

$$(1) \quad \frac{dx}{dt} = Ax + Bu$$

where  $u = (u_1, u_2, \dots, u_r)$  is a control vector that must be chosen for each  $t$  from a convex compact set  $U(t)$ . The initial conditions at  $t = 0$  are

$$x^0 = (0, \xi_1^0, \xi_2^0, \dots, \xi_n^0), \text{ (fixed) } .$$

The terminal conditions at  $t = T$  is obtained by setting

$$(2) \quad x^T = \bar{x}^T + x x_0$$

where

$$\bar{x}^T = (0, \xi_1^T, \xi_2^T, \dots, \xi_n^T), \text{ (fixed)}$$

$$E_0 = (1, 0, 0, \dots, 0)$$

and by requiring that  $u = u(t)$  to be chosen such that

$$(3) \quad Z \text{ is minimum} \quad .$$

As given in Chapter III of the book "Mathematical Theory of Optimal Control Processes" by Pontryagin, Boltyashii, Gekhtel'de, Mischenko, the final state may be written in the form

$$(4) \quad -ZE_0 + \int_0^T P_{T-t} B u(t) dt = b$$

where  $b = \bar{x}^T - P_T \bar{x}^0$  is a known vector, and  $P_t = e^{tA}$  matrix that may be conveniently computed as a function of  $t$ . For example, for the case of real distinct characteristic roots  $\lambda_i$  of  $A$ :

$$(5) \quad P_t = e^{tA} \sum_{i=1}^n M_i e^{\lambda_i t}$$

where  $M_i$  are square matrices independent of  $t$ . The latter formula for the  $M_i$  is developed in "An Introduction to the Application of Dynamic Programming to Linear Control Systems" by F. T. Smith in RAND Report RM-3526-PR, February 1963.

We may formally write (4) as a generalized linear program.

PROBLEM: Find  $\text{Min } Z, \mu \geq 0$  such that

$$(6) \quad \begin{aligned} -ZE_0 + Y\mu &= b \\ \mu &= 1 \end{aligned}$$

where  $Y$  may be freely chosen from the convex set defined by

$$(7) \quad Y \geq \int_0^T P_{T-t} B u(t) dt$$

for all possible choices of  $u(t) \in U(t)$ .

The method for solving the generalized linear program described earlier for convex programming can be applied. For brevity, we omit the question of how to obtain the initializing basic set, except to say it is the analog of the phase I procedure of the ordinary simplex method.

As soon as  $\pi = \pi^k$  is determined for iteration  $k$  we seek a solution of the sub-problem such that the inner product

$$\begin{aligned} \text{Min } \pi^k Y &= \text{Min } \pi^k \int_0^T P_{T-t} B u(t) dt \\ &= \int_0^T [\text{Min } (\pi^k P_{T-t} B u(t))] dt \quad \text{for } u(t) \in U(t). \end{aligned}$$

It is important to note that for each  $t$ ,  $\pi^k P_{T-t} B$  is some known vector  $c^t$ .

Thus, for each  $t$  we must solve:

SUB-PROBLEM: Find  $\text{Min } c^t u$  for  $u \in U(t)$ .

If  $U(t)$  is a polyhedral set, the sub-problem is simply a linear program. If

$U(t)$  is the same for all  $t$ , then the linear programs are the same for all  $t$  except for the varying objective forms  $c^t u$ . Interesting enough, the sub-programs turn out to be the same as what Pontryagin obtains using his maximal principle.

A control problem is an example of a dynamic system which, if it is treated by the straightforward procedure of discretizing the time, would lead to a large-scale computational problem. The generalized linear program (or decomposition principle approach), however, provides us with a procedure which does not require the discretizing of the time interval.

The last twenty years have been marked by the accelerated trend toward automation. Many believe that not only simple control processes, but soon the more complex control processes will be mechanized. If so, whether we like it or not, decisions will be made for us by machines. Whether or not they will be good decisions will depend on how cleverly we have instructed the machines. This in turn will depend heavily on how clever we have been in developing solution techniques for solving large-scale systems.

To this end, we have sketched several ideas: (1) taking advantage of the low density of the non-zero coefficients in the original matrix, (2) finding a compact inverse representation of the basis using the simplex method, and (3) making use of the generalized linear program or decomposition principle approach. We illustrated the latter on a linear-control problem and found that it led to the maximal principle with the added bonus, however, that it can be used to constructively converge to an optimal solution.

## LINEAR CONTROL PROCESSES AND MATHEMATICAL PROGRAMMING\*

GEORGE B. DANTZIG†

Linear control process defined [8], [14]. We shall consider an "object" defined by its  $n + 1$  coordinates  $X = (x_0, x_1, \dots, x_n)$ , whose "motion" described as a function of a parameter, "time" ( $t$ ), can be written as a linear system of differential equations

$$(1) \quad \frac{dX}{dt} = A'X + B'u,$$

where  $A'$ ,  $B'$  are known matrices that may depend on  $t$  and

$$u = (u_1, u_2, \dots, u_m)$$

is a control vector that must be chosen from a convex set,  $u \in U(t)$  for every  $0 \leq t \leq T$ . The time period  $0 \leq t \leq T$  is fixed and known in advance. The coordinate  $x_0 = x_0(t)$  represents the "cost" of moving the object from its initial position to  $x_0(t)$ . For this purpose it may be assumed that  $x_0(0) = 0$ . Defining

$$(2) \quad \hat{X} = (0, x_1, x_2, \dots, x_n),$$

the object is required to start somewhere in a convex domain  $\hat{X}(0) \in S_0$  and to terminate at  $t = T$  somewhere on another convex domain  $\hat{X}(T) \in S_T$ .

*Problem.* Find  $u \in U(t)$  and boundary values  $\hat{X}(0) \in S_0$ ,  $\hat{X}(T) \in S_T$ , such that  $x_0(T)$  is minimized.

Assuming  $u \in U(t)$  is known, the system of differential equations can be integrated to yield an expression for  $X(T)$  in terms of  $X(0)$  and  $u \in U(t)$ . This is true in general but will be illustrated for the case when  $A'$  and  $B'$  do not depend on  $t$ ; in this case

$$(3) \quad X(T) = e^{TA}X(0) + \int_0^T e^{(T-t)A}Bu(t) dt,$$

where  $u(t) \in U(t)$  is a convex set and where we assume the integral exists whatever be the choice of the  $u(t) \in U(t)$  for  $0 \leq t \leq T$ .

**Generalized linear program [2].** Our general objective is to illustrate

\* Received by the editors January 12, 1965, and in revised form March 25, 1965. Presented at the First International Conference on Programming and Control, held at the United States Air Force Academy, Colorado, April 15, 1965.

† Operations Research Center, University of California, Berkeley, California. This research has been partially supported by the National Science Foundation under Grant GP-2633 with the University of California.

how *mathematical programming* and, in particular, how the *decomposition principle* in the form of the *generalized linear program* can be applied to this class of problems. An elegant constructive theory emerges, [10], [11], [12], [13].

A *generalized linear program* differs from a standard linear program in that the vector of coefficients, say  $P$ , associated with any variable  $\mu$  need not be constant but can be selected from a convex set  $C$ . For example:

*Problem.* Find  $\max \lambda, \mu \geq 0$  such that

$$(4) \quad U_0 \lambda + P \mu = Q_0, \quad \mu = 1,$$

where  $U_0, Q_0$  are specified vectors and  $P \in C$  convex.

It is assumed that the elements of  $C$  are only known implicitly (for example, as some solution to a linear program) but that particular choices of  $P$  can be easily obtained which minimize any given linear form in the components of  $P$ .

The method of solution assumes we have initially<sup>1</sup> on hand  $m$  particular choices  $P_i \in C$  with the property that

$$(5) \quad \begin{aligned} U_0 \lambda + P_1 \mu_1 + P_2 \mu_2 + \cdots + P_m \mu_m &= Q_0, \\ \mu_1 + \mu_2 + \cdots + \mu_m &= 1, \end{aligned}$$

has a unique "feasible" solution; that is to say,  $\lambda = \lambda^0, \mu_i = \mu_i^0 \geq 0$  and the matrix

$$(6) \quad B^0 = \begin{bmatrix} U_0 & P_1 & \cdots & P_m \\ 0 & 1 & \cdots & 1 \end{bmatrix}$$

is nonsingular (i.e., the columns of  $B^0$  form a basis). Because  $P_i \in C$ , the vector  $P^0 = \sum P_i \mu_i^0$  constitutes a solution  $P = P^0$  for (4) except that  $\lambda = \lambda^0$  may not yield the maximal  $\lambda$ .

To test whether or not  $P^0$  is an optimal solution one determines a row vector  $\bar{\pi} = \bar{\pi}^0$  such that

$$(7) \quad \bar{\pi}^0 B^0 = (1, 0, \cdots, 0),$$

and then a value  $\delta$  and a vector  $P_{m+1} \in C$  such that

$$(8) \quad \delta = \bar{\pi}^0 P_{m+1} = \min_{P \in C} \bar{\pi}^0 P,$$

where we denote

$$(9) \quad \bar{P} = \begin{bmatrix} P \\ 1 \end{bmatrix}.$$

If it turns out that  $\delta = 0$ , then  $P = P^0$  is an optimal solution.

<sup>1</sup> This is not a restrictive assumption since there is an analogous method for obtaining such a starting solution, see [2].

If  $P^0$  is not optimal, system (5) is augmented by  $P_{m+1}$ . After one or several iterations  $k$  the augmented system takes the form of a linear program:

*Problem.* Find  $\max \lambda, \mu_i \geq 0$ ,

$$(10) \quad U_0 \lambda + \sum_1^{m+k} P_i \mu_i = Q_0, \quad \sum_1^{m+k} \mu_i = 1.$$

Letting  $B^k$  denote the basis associated with an optimal basic feasible solution  $\mu_i = \mu_i^k$  to (10),  $\pi^k$  is defined analogous to (7) and  $\delta^{k+1}$  and  $P_{m+k+1}$  analogous to (8). If it turns out that  $\delta = 0$ , the solution

$$(11) \quad P^k = \sum_1^{m+k} P_i \mu_i^k$$

is optimal. If not the system is augmented by  $P_{m+k+1}$  and the iterative process is repeated.

It is known under certain general conditions, such as  $C$  bounded and the initial solution nondegenerate (i.e.,  $\mu_i^0 > 0$ ), that  $\bar{\pi}^k \rightarrow \bar{\pi}^*$  and  $P^k \rightarrow P^*$  on some subsequence  $k$  and that  $P = P^*$  is optimal. The two fundamental properties of  $\bar{\pi}^*$  are

$$(12) \quad \bar{\pi}^* \neq 0 \quad \text{and} \quad \bar{\pi}^* \bar{P} \geq \bar{\pi}^* P^* = 0 \quad \text{for all } P \in C.$$

The entire process can be considered as constructive providing it is not difficult to compute the various  $P_{m+k+1}$  from (8) with  $\bar{\pi} = \bar{\pi}^{m+k}$ . For example, if  $C$  is a parallelepiped or more generally a convex polyhedral set, then  $\min \bar{\pi} \bar{P}$  constitutes the minimization of a linear form with known coefficients  $\bar{\pi} = \bar{\pi}^{m+k}$  subject to linear inequality constraints in the unknown components of  $\bar{P}$ , i.e., a linear program. In this case the iterative process terminates in a finite number of steps and  $P_{m+k}$  constitute extreme solutions from it. In all cases an estimate is available on how close the  $k$ th solution is to an optimal value of  $\lambda$ .

**Application of the generalized program to the linear control process.** Let us denote

$$(13) \quad P = \int_0^T e^{(T-t)A} B u(t) dt,$$

and note that  $P$  is an element of a convex set  $C$ , generated by choosing all possible  $u(t) \in U(t)$ . We specify that  $U_0 = (1, 0, \dots, 0)$ , and denote by  $\lambda = -X_0(T)$ , where  $X_0(T)$  is the coordinate of  $X(T)$  to be minimized. Then

$$(14) \quad X(T) = -U_0 \lambda + \bar{X}(T).$$

We further define  $Q_0$  by

$$(15) \quad \hat{X}(T) = e^{TA}X(0) + Q_0.$$

Substitution of these into (3) formally converts<sup>2</sup> the integrated form of the control problem into a generalized linear program (4).

Each cycle of the iterative process yields a known row vector, which we partition

$$(16) \quad \bar{\pi}^{k+1} = [\pi, \theta],$$

where  $\pi$  represents its first  $n + 1$  components corresponding to  $P$  and  $\theta$  its last component. Since  $\pi$  is known, our choice for  $P_{m+k+1}$  is

$$(17) \quad \pi P_{m+k+1} = \min \left\{ \int_0^T \pi e^{(T-t)A} B u(t) dt \right\} = \int_0^T \left\{ \min_{u \in U(t)} e^{(T-t)A} B u(t) \right\} dt,$$

where clearly the minimum is obtained when, in (17), the integrand for each  $t$  is selected to be minimum.

Note that

$$(18) \quad \phi_{t,\pi} = \pi e^{(T-t)A} B$$

is a row vector that can be computed for each  $t$ . For example,  $\phi_{t,\pi}$  can be represented by a finite sum of vectors whose weights depend on  $t$  and the eigenvalues of  $A$ . The new extremal solution  $P_{m+k+1}$  is obtained by choosing the control which minimizes the linear form in  $u$  for each  $t$ ; i.e., find

$$(19) \quad \min (\phi_{t,\pi} u), \quad u \in U(t).$$

For example, if  $U(t)$  is a polyhedral set then (19) is a linear program. If  $U(t)$  is the same for all  $t$ , then only the objective form,  $\phi_{t,\pi} u$ , varies for different  $t$ ; except for the objective form the linear programs are the same for all  $t$ .

If optimal  $\pi^*$  is used, then the optimal control  $u$  (except for a set of measure zero) satisfies

$$(20) \quad \min \{\phi^*(t)u\}, \quad u \in U(t),$$

where  $\phi^*(t) = \pi^* e^{(T-t)A} B$ . Pontryagin refers to this as the *maximal principle*. It is, as we have just shown, also a consequence of the decomposition principle of linear programming.

**Conclusion.** In our approach the general control obtained for each cycle is a linear combination of exactly  $n + 1$  special controls obtained by mini-

<sup>2</sup> Actually  $Q_0$  is not given but is an element of a convex set. To simplify the discussion which follows we assume  $Q_0$  is a fixed vector.

mizing for each  $t$ , the linear expression (19) in  $u$  for  $n + 1$  choices of  $\pi$ . These special controls may be referred to as extreme controls. The latter each in themselves do not maintain feasibility, that is to say, guarantee that the object will move from  $\bar{X}(0)$  to  $\bar{X}(T)$ . Each new linear combination of these special controls will, however, generate a new feasible control with a lower value<sup>1</sup> for the total cost  $X_0(T)$ . Under the conditions stated this iterative process is known to converge.

## REFERENCES

- [1] R. BELLMAN, *Adaptive Control Processes: A Guided Tour*, Princeton University Press, Princeton, 1961.
- [2] G. B. DANTZIG, *Linear Programming and Extensions*, Princeton University Press, Princeton, 1963.
- [3] A. F. FILIPPOV, *On certain questions in the theory of optimal control*, English transl., this Journal, 1(1962), pp. 76-84.
- [4] H. HALKIN, *On the necessary conditions for optimal control of nonlinear systems*, J. Analyse Math., 12 (1964), pp. 1-52.
- [5] J. P. LASALLE, *The time optimal control problem*, Contributions to the Theory of Nonlinear Oscillations, vol. V, Princeton University Press, Princeton, 1958.
- [6] J. P. LASALLE AND S. LEFSCHETZ, *Stability by Liapunov's Direct Method*, Academic Press, New York, 1961.
- [7] G. LEITMANN, ed., *Optimization Techniques*, Academic Press, New York, 1962.
- [8] L. W. NEUSTADT, *Discrete time optimal control systems*, Nonlinear Differential Equations and Nonlinear Mechanics, J. P. LaSalle and S. Lefschetz, eds., Academic Press, New York, 1963.
- [9] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE, AND E. F. MISHCHENKO, *The Mathematical Theory of Optimum Processes*, Interscience, New York, 1962.
- [10] B. H. WHALEN, *Linear programming for optimal control*, Ph.D. dissertation, University of California, Berkeley, 1962.
- [11] ———, *On linear programming and optimal control*, Correspondence to IRE Trans. on Automatic Control, AC-7 (1962), p. 46.
- [12] R. M. VAN SLYKE, *Mathematical programming*, Ph.D. dissertation, University of California, Berkeley, 1965.
- [13] L. A. ZADEH, *A note on linear programming and optimal control*, Correspondence to IRE Trans. on Automatic Control, AC-7 (1962), p. 46.
- [14] L. A. ZADEH AND C. A. DESOER, *Linear System Theory, The State-Space Approach*, McGraw-Hill, New York, 1963.

---

<sup>1</sup> If basic solution is nondegenerate.

ALL SHORTEST ROUTES IN A GRAPH

BY

GEORGE B. DANTZIG

TECHNICAL REPORT NO. 66-3

November 1966

Operations Research House  
Stanford University  
Stanford, California

Research of G. B. Dantzig partially supported by Office of Naval Research, Contract ONR-N-00014-67-A-0112-0011, U. S. Atomic Energy Commission, Contract No. AT(04-3)-326 PA #18, and National Science Foundation Grant GP 6431; reproduction in whole or in part for any purpose of the United States Government is permitted.

# ALL SHORTEST ROUTES IN A GRAPH

by

George B. Dantzig

A shortest route is sought between every pair of nodes  $(i, j)$  in a graph when directed arc distances  $a_{ij}$  are given, where the values of  $a_{ij}$  may be positive, negative, or zero except  $a_{ii} = 0$ . If the graph is incomplete so that an arc  $(i, j)$  is missing, the value of  $a_{ij} = \infty$ . This problem (as is well known) includes the travelling salesman problem since the route for  $(i, i)$  is a cycle and one can solve a travelling salesman problem with distances  $d_{ij} > 0$  by finding a minimum cycle in a graph  $[a_{ij} = d_{ij} - K]$  where  $K > \sum_i \sum_j d_{ij}$ . Our objective, therefore is more modest, it is to find a negative cycle in a graph if one exists, if none then to find all the shortest routes.

The procedure is inductive and was stimulated by a remark of Ralph Gomory's that an inductive approach was probably as efficient as any other. It is not certain, however, whether this procedure has appeared elsewhere in the literature and so is presented here.

It is shown that  $n(n-1)^2$  additions and an equal number of comparisons are required to solve an  $n$  node problem. This number can be reduced to  $n(n-1)(n-2)$  if negative cycles are known not to exist. This method is therefore as efficient as the best result known, that of Murchland [3].

It is similar to many proposed schemes in that entries  $a_{ij}$  in the matrix are replaced by  $a_{ik} + a_{kj}$  if the latter sum is smaller for some choice of  $k$ . After replacement the new matrix is operated upon in the same way until no improvement can be found. The various methods differ only in the rules for scanning the various  $(i, j)$  and  $k$ . In order to keep track of the routes as well as their values, it is also necessary to record for each  $(i, j)$  either the first arc of the minimum route from  $i$  to  $j$  or the last arc. With this information it is easy to generate all the arcs along the route. Aside from the efficiency, the second advantage of the method is the simplicity of the proof of its validity.

Assume for nodes  $1, 2, \dots, k-1$  that optimal distances  $\bar{a}_{ij}$  are given, we wish to determine optimal distances  $a_{ij}^*$  for nodes  $1, 2, \dots, k$ . We shall show that

For  $l = 1, \dots, (k-1)$

$$(1) \quad a_{kl}^* = \min_j (a_{kj} + \bar{a}_{jl}) \quad j = 1, 2, \dots, k-1$$

$$(2) \quad a_{lk}^* = \min_j (a_{jk} + \bar{a}_{lj})$$

$$(3) \quad a_{kk}^* = \min_j [0, a_{kj}^* + a_{jk}^*]$$

For  $(i = 1, \dots, k-1)$  and  $(j = 1, \dots, k-1)$

$$(4) \quad a_{ij}^* = \min [ \bar{a}_{ij}, a_{ik}^* + a_{kj}^* ]$$

The inductive procedure begins with  $\bar{a}_{11} = 0$  and stops if at any time a diagonal value  $a_{11}^* < 0$  appears in which case a negative

cycle has been obtained; or if step  $k = n$  has been completed.

Proof: (1) states that a minimum route from  $k$  to  $l$  starts with some arc  $a_{kj}$  followed by a minimum route from  $j$  to  $l$  that does not go through  $k$ . Hence the minimum of these alternative routes is the one desired.

Formula (2) is the same idea except the alternative routes are defined by the last arc  $a_{jk}$  of the route and the best route from  $l$  to  $j$  that does not go through  $k$ .

Formula (3) states that either  $a_{kk} = 0$  is the best route from  $k$  to  $k$  or there is a negative cycle consisting of going along some best route from  $k$  to  $i$  and then  $i$  back to  $k$ .

Formula (4) states that either the best route from  $(i, j)$  does not go through  $k$  (and has value  $\bar{a}_{ij}$ ) or does go through  $k$  (and has value  $a_{ik}^* + a_{kj}^*$ ).

The count on additions is

$$C = \sum_{k=1}^n [(k-1)(k-2) + (k-1)(k-2) + (k-1) + (k-1)^2]$$

where the four terms are the count of (1), (2), (3), (4)

respectively. Note that we omitted from the count (for example)

the addition  $a_{kl}^* + \bar{a}_{ll}$  because  $\bar{a}_{ll}$  is known to be

zero. In the case that negative cycles are known not to exist, the third term may be dropped and the last term reduced by  $(k-1)$  since the diagonal  $a_{ii}^* = 0$ . In the latter case, the count is

$C = n(n-1)(n-2)$  additions and an equal number of comparisons.

## REFERENCES

- [1] Farbey, B. A., Land, A. H., Murchland, J. D., "The Cascade Algorithm for Finding the Minimum Distances on a Graph." Transport Network Theory Unit, London School of Economics, October 1966.
- [2] Dantzig, George, B., "On the Shortest Route Through a Network," Management Science, Vol. 6, No. 2, January, 1960.
- [3] Dantzig, George, B., Linear Programming and Extensions, Princeton University Press, Princeton, N. J., 1963, pages 361-66.
- [4] Hu, T. C., "Revised Matrix Algorithms for Shortest Paths," IBM Watson Research Center, Research Paper, RC-1478, September 28, 1965.
- [5] Hu, T. C., "A Decomposition Algorithm for Shortest Paths in a Network," IBM Watson Research Center, Research Paper, RC-1562, February 4, 1966.
- [6] Murchland, J. D., "An Inductive Matrix Method for Finding All Shortest Paths in a Directed Graph," LSE-TNT-25, March 29, 1966.
- [7] Murchland, J. D., "A New Method for Finding All Elementary Paths in a Complete Directed Graph," LSE-INT-22, October 26, 1965.
- [8] Murchland, J. D., "An Algol Procedure for all Shortest Paths in a Symmetric Graph," LSE-INT-35, March 7, 1966.
- [9] Murchland, J. D., "An Inductive Matrix Method for Finding all Shortest Paths in a Directed Graph," LSE-INT-25, March 29, 1966.
- [10] Murchland, J. D., "The Extension of the Cascade Algorithm to Large Graphs," LSE-INT-20, September, 1966 (Revised)
- [11] Murchland, J. D., "Bibliography of the Shortest Route Problem," LSE-TNT-6, June 1966. (Revision)
- [12] Murchland, J. D., "A New Method for Finding all Elementary Paths in a Complete Directed Graph," LSE-TNT-22, October 26, 1965.

ALL SHORTEST ROUTES FROM A FIXED ORIGIN IN A GRAPH

BY

G. B. Dantzig\*, W. O. Blattner,\*\* and M.R. Rao\*\*

TECHNICAL REPORT NO. 66-2

November 1966

\*Operations Research House  
Stanford University  
Stanford, California

\*\*United States Steel Corporation

Research Of G. B. Dantzig partially supported by Office of Naval Research, Contract ONR-N-00014-67-A-0112-0011, U. S. Atomic Energy Commission, Contract No. AT(04-3)-326 PA #18, and National Science Foundation Grant GP 6431; reproduction in whole or in part for any purpose of the United States Government is permitted.

# ALL SHORTEST ROUTES FROM A FIXED ORIGIN IN A GRAPH

by

G. B. Dantzig\*, W. Blattner\*\* and M. R. Rao\*\*

A shortest route is sought between a fixed origin node  $i = 0$  to  $n$  other nodes in a graph when directed arc distances  $c_{ij}$  are given and the values of  $c_{ij}$  may be positive, negative, or zero  $i \neq j$ . No values  $c_{ij}$  are specified unless there is an arc from  $i$  to  $j$ . This problem (as is well known) includes the travelling salesman problem with distances  $d_{ij} > 0$  because one can set  $[c_{ij} = d_{ij} - K]$  where  $K > \sum_i \sum_j d_{ij}$  and look for a minimum route from 0 back to itself. Therefore our objective will be more modest: To find a negative cycle in a graph if one exists or if none exists then to find all the shortest paths from the origin.

The method is inductive. On step  $k$ , there is a set  $S_k$  consisting of the origin and  $k - 1$  other nodes. Restricting arcs to those that belong to the subgraph of  $S_k$ , the minimum distances from the origin along these arcs to nodes  $i \in S_k$  are assumed known and have value  $\pi_i$ . It is also assumed that no negative cycles exist in the subgraph of  $S_k$ . It follows that

$$(1) \quad \pi_i + c_{ij} \geq \pi_j \quad \text{for all } i \in S_k, j \in S_k.$$

---

\* Stanford University

\*\* U. S. Steel

Theorem 1: Let  $D_{ij}$  denote the length of the shortest route from  $i$  to  $j$  along arcs of the subgraph of  $S_k$  containing no negative cycles and let (1) hold, then

$$(2) \quad D_{ij} \geq \Pi_j - \Pi_i$$

Proof: Let the sequence  $(i; i_1, i_2, \dots, i_\lambda; j)$  denote the nodes along a minimum route from  $i$  to  $j$  in  $S_k$ , then by (1),

$$\Pi_i + c_{ii_1} \geq \Pi_{i_1}, \quad \Pi_{i_1} + c_{i_1 i_2} \geq \Pi_{i_2}, \dots, \quad \Pi_{i_\lambda} + c_{i_\lambda j} \geq \Pi_j.$$

Adding these inequalities together yields the desired relation.

Assuming now that we know the minimal distances  $\Pi_i$  for  $S_k$ , we wish to augment  $S_k$  by including a node  $q \notin S_k$ . We denote  $S_{k+1} = \{S_k, q\}$  and wish to determine minimal distances  $\Pi_i^*$  from the origin along arcs of the subgraph of  $S_{k+1}$  to nodes  $i \in S_{k+1}$ . The theorem below permits us to determine  $\Pi_q^*$  immediately.

Theorem 2: Let  $q \notin S_k$ , and  $S_{k+1} = \{S_k, q\}$  then a shortest route from  $0$  to  $q$  in  $S_{k+1}$  has as last arc of the route  $(p, q)$  where  $p \in S_k$  satisfies

$$(3) \quad \Pi_p + c_{pq} = \min_{i \in S_k} (\Pi_i + c_{iq})$$

and  $\Pi_q^* = \Pi_p + c_{pq}$  is the minimum distance from the origin to  $q$  in  $S_{k+1}$ .

Proof: Suppose false and a shorter route is via  $\bar{p} \in S_k$ , then

$$\Pi_{\bar{p}} + c_{\bar{p}q} < \Pi_p + c_{pq}$$

contradicting (3). This theorem is true even if  $S_k$  has negative cycles. The  $\Pi_q^*$  and  $\Pi_1$  would then represent the shortest distance without cycles from the origin.

Knowing  $\Pi_q^*$ , Theorem (4) below may now be applied to determine for another node  $l \in S_{k+1}$ , its minimum distance  $\Pi_l^*$  from the origin along arcs of the subgraph of  $S_{k+1}$ . Knowing  $\Pi_q^*$  and  $\Pi_l^*$  we reapply Theorem (4) again and again, each time finding a least distance for another node in  $S_{k+1}$ . This is done until all nodes are exhausted in  $S_{k+1}$  or the optimality condition  $\delta_{ij} \geq 0$  of Theorem 3 below is satisfied in which case the remaining  $\Pi_1$  values are also optimal for  $S_{k+1}$ , or the negative cycle condition of Theorem 5 is satisfied.

Theorem 3: Let  $T$  be any subset of nodes  $i$  whose minimum distance  $\Pi_1^*$  from the origin along routes in the subgraph of  $S_{k+1}$  is known,  
let  $q \in T$ ; let  $S_k$  and  $T$  contain no negative cycles; let

$$(4) \quad \delta_{1j} = \Pi_1^* + c_{1j} - \Pi_j \quad i \in T, j \notin T$$

then, if

$$(5) \quad \delta_{1j} \geq 0 \quad \text{for all } i \in T, j \notin T$$

the minimum distance for all remaining nodes is

$$(6) \quad \Pi_j^* = \Pi_j \quad \text{for all } j \notin T$$

This theorem is true even if  $T$  contains negative cycles but requires a different proof.

Proof: The conditions for optimality in  $S_{k+1}$  analogous to (1) are:

$$(7) \quad \delta_{ij} = \pi_i^* + c_{ij} - \pi_j \geq 0 \quad i \in T, j \notin T$$

$$\pi_i + c_{ij} - \pi_j \geq 0 \quad i \notin T, j \notin T$$

$$\pi_i^* + c_{ij} - \pi_j^* \geq 0 \quad i \in T, j \in T$$

$$\pi_i + c_{ij} - \pi_j^* \geq 0 \quad i \notin T, j \in T$$

The first of these holds by hypothesis (5), the second by (1), the third by hypothesis that the  $T$  set is optimal in  $S_{k+1}$  (and there are no negative cycles in  $T$ ); finally the fourth because  $\pi_j^* \leq \pi_j$  and (1) holds.

On the other hand if the optimality conditions  $\delta_{ij} \geq 0$  of Theorem 3 does not hold for all  $i \in T, j \notin T$ , then  $\delta_{t\ell} = \min \delta_{ij} < 0$  holds for some  $t \in T$  and  $\ell \notin T$ . It will be shown in Theorem 4, that the minimum distance from the origin along arcs of the subgraph of  $S_{k+1}$  to node  $\ell$  is given by  $\pi_\ell^* = \pi_t + \delta_{t\ell}$ . Thus Theorem 4 may be reapplied until there are no longer any nodes in  $S_{k+1}$  not in  $T$  or condition (5) holds, or a negative cycle is detected, but we will speak more about this later in Theorem 5.

Theorem 4: Let  $S_k$  and  $T$  contain no negative cycles where  $T$  is any subset of nodes  $i$  whose minimum distances from the origin in

$S_{k+1}$  is  $\Pi_1^*$ . If for some  $t \in T$ ,  $l \notin T$

$$(8) \quad \delta_t = \min_{j \in T} \delta_{tj} < 0 \quad t \in T, j \notin T$$

then

$$(9) \quad \Pi_l^* = \Pi_l + \delta_{tl} = \Pi_t^* + c_{tl}$$

is the minimal distance from the origin along arcs in the subgraph of  $S_{k+1}$  to node  $l$ .<sup>1)</sup>

Proof: On the contrary, if there is a shorter route to  $l$ , then this route must include the node  $q$  and perhaps some other nodes of  $T$  (otherwise  $\Pi_l$  would be minimum but we know  $\Pi_l^* < \Pi_l$  by (8) and (9). Along this shorter route let  $(\bar{t}, \bar{l})$  be the last arc such that  $\bar{t} \in T$ ,  $\bar{l} \notin T$ . Then the distance along the route from  $\bar{l}$  to  $l$ , may be denoted by  $D_{\bar{l}l}$  (see Theorem 1) because the nodes from  $\bar{l}$  to  $l$  are all elements of  $S_k$ . By Theorem (1)

$$(10) \quad D_{\bar{l}l} \geq \Pi_l - \Pi_{\bar{l}}$$

On the other hand by virtue of the assumed shorter route through  $\bar{t}, \bar{l}$

$$(11) \quad \Pi_{\bar{t}}^* + c_{\bar{t}\bar{l}} + D_{\bar{l}l} < \Pi_t^* + c_{tl}$$

---

<sup>1)</sup> This theorem also holds if  $T$  contains negative cycles and  $\Pi_1^*$  are the shortest distances from the origin along routes without cycles.

Subtracting (10) from (11) and rearranging

$$\pi_t^* + c_{t\bar{l}} - \pi_{\bar{l}} < \pi_t^* + c_{tl} - \pi_l$$

or  $\delta_{t\bar{l}} < \delta_{tl}$  by (4) which contradicts hypothesis (8) of Theorem 4.

Theorem 5: If  $S_k, T$  contain no negative cycles and the shortest distance from the origin in  $S_{k+1}$  for  $i \in T$  is  $\pi_i^* < \pi_i$  and  $T$  is augmented to  $T^* = \{T, \bar{l}\}$  where  $\bar{l}$  is as defined in Theorem 4, then a necessary and sufficient condition that  $T^*$  contain a negative cycle is

$$(12) \quad \pi_{\bar{l}}^* + c_{\bar{l}q} - \pi_q^* \delta_{\bar{l}q} < 0$$

Proof: Since  $\pi_i^* < \pi_i$  holds the optimal route from the origin to  $\bar{l}$  in  $S_{k+1}$  passes through  $q$ . If (12) holds, then the cycle consisting of the optimal route from  $q$  to  $\bar{l}$  and then arc  $(\bar{l}, q)$  has negative length. This may be seen by summing the relations  $\pi_i^* + c_{ij} = \pi_j^*$  along the route from  $q$  to  $\bar{l}$  and then adding it to (12). If, on the other hand, (12) does not hold, then we will show that  $\pi_i^* + c_{ij} \geq \pi_j^*$  for all  $i \in T^*, j \in T^*$  which implies that no negative cycle in  $T^*$  exists (as one can see by summing such relations over the arcs of a cycle.)

We need now only rule out for some  $i$  and  $j \neq q$  that  $\pi_{i_1}^* + c_{i_1 j_1} < \pi_{j_1}^*$ . This would mean we could lower the value of  $\pi_{j_1}^*$  by making  $i_1$  the node that precedes  $j_1$  along the optimal route instead of some  $i_1$ . This deletion of the arc  $(i_1 j_1)$  from

the tree<sup>2)</sup> of optimal routes and entering the arc  $(i, j_0)$  into the tree either would provide a shorter route to  $j_0$  or it would cause a cycle to form which (by an earlier argument) is negative.

However neither is possible because the former implies a shorter route to  $j_0$  (because  $\Pi_{j_0}^*$  was lowered) while the latter implies a negative cycle not involving  $q$ . The cycle cannot involve  $q$  because all shortest routes  $i \in T^*$  from the origin pass through  $q$  and there are no directed arcs into  $q$  along the tree of optimal routes in  $T^*$ . But a negative cycle in  $S_k$  is contrary to assumption.

Thus a negative cycle will always be found if there is one by (12). If one is found the inductive process terminates.

The following theorem due to M. Sakarovitch (verbal communication) permits one to find the minimal distance in  $S_{k+1}$  to several nodes at once.

Theorem 6 (Sakarovitch): Let  $L$  be the nodes in the tree of optimal routes in  $S_k$  which are successors<sup>3)</sup> of  $l$  as defined in Theorem 4, then

$$(13) \quad \Pi_i^* = \Pi_l + \delta_{tl} \quad \text{for } i \in L.$$

<sup>2)</sup> Note: If there are no negative cycles in  $S_k$  and  $T$  in  $S_{k+1}$  there is a tree of optimal routes to  $i \in T$  branching out from the origin; also the added arc  $(t, l)$  with  $t \in T$ ,  $l \notin T$  still yields a tree of shortest routes without cycles in  $i \in T^*$ .

<sup>3)</sup> The tree of optimal routes from the origin forms a partially ordered set. The "successors" of  $l$  are those nodes reached through  $l$ .

Proof: One notes first that the distance  $\Pi_1 + \delta_{t\ell}$  can be realized by first going along the optimal route to  $\ell$  and then along the former route from  $\ell$  to  $i \in L$ . Now assume on the contrary that there is a better route to  $i$ . As in proof of Theorem 4, let  $\bar{t}\bar{\ell}$  be the last arc of a better route such that  $\bar{t} \in T$  and  $\bar{\ell} \notin T$ , then  $\Pi_{\bar{t}}^* + c_{\bar{t}\bar{\ell}} + D_{\bar{\ell}i} < \Pi_1 + \delta_{t\ell}$ . Subtracting  $D_{\bar{\ell}i} \geq \Pi_{1\ell} - \Pi_{\bar{\ell}}$ , yields  $\delta_{\bar{t}\bar{\ell}} < \delta_{t\ell}$  contrary to (8).

For completeness we give the following well known theorem, [3].

Theorem 7: If  $c_{ij} \geq 0$  and  $\Pi_1$  of  $S_k$  are known to be the minimal distances from the origin for the  $k$  nodes of  $S_k$  using arcs of the full  $n$ -node problem, then  $\Pi_q = \Pi_p + c_{pq}$  is the minimal distance for  $q \notin S_k$  where

$$(14) \quad \Pi_p + c_{pq} = \text{Min} (\Pi_1 + c_{1j}) \quad , p \in S_k$$

$$i \in S_k$$

$$j \notin S_k$$

Proof: If not, then  $q$  is reached via some shorter route that has nodes in common with  $S_k$  (since  $S_k$  includes the origin). Let  $(\bar{t}, \bar{q})$  be the last arc on the shorter route with  $\bar{t} \in S_k$  and  $\bar{q} \notin S_k$ , then

$$(15) \quad \Pi_{\bar{t}} + c_{\bar{t}\bar{q}} + (\text{min distance } \bar{q} \text{ to } q) < \Pi_p + c_{pq}$$

but this relation contradicts (14) because minimum distance from  $\bar{q}$  to  $q$  is non-negative when  $c_{ij} \geq 0$ .

We are now in a position to give a count on the number of

additions. Associated with each set of additions such as for (14) is the same number of comparisons (or possibly one less). In the case  $c_{ij} \geq 0$ , the same sums occur in  $S_k$  and  $S_{k+1}$  for the same  $(i, j)$ . Since at step  $k+1$  we do not need to consider the arcs back to  $S_k$ , the total additions do not exceed the total number of arcs. We will denote this total by  $A$ . The procedure is to sort the  $\Pi_i + c_{ij}$  values as generated from low to high. Let the lowest sum on this list be  $\Pi_i + c_{ij}$ . This sum on the list is deleted if  $\Pi_j$  has previously been determined; if not then  $\Pi_j = \Pi_i + c_{ij}$ . Next the sums  $\Pi_j + c_{jk}$  are computed for all arcs  $(j, k)$  and made part of the sorted list. The process is then repeated. Sorting requires effort, however, and so that the two theorems that follow are misleading.

Theorem 8: If all distances  $c_{ij} \geq 0$ , then the number of additions using formula (14) does not exceed  $A$ , the number of arcs.

Theorem 9: The number of additions in the general case, when formula (3) and (8) is used does not exceed

$$(16) \quad A + nf_1 + (n-1)f_2 + \dots + f_n$$

where  $n$  is the number of nodes,  $f_k$  is number of arcs directed forward from the  $k$ -th node to enter the induction.

This suggests preordering from low to high the nodes by the number of their forward arcs. If this is done, the bound reduces to

$$(17) \quad A + nf_1 + (n-1)f_2 + \dots + f_n \leq (n+3)A/2$$

## REFERENCES

- [1] Dantzig, G. B., Blattner, W. C., and Rao, M. R., "Finding a Cycle in a Graph with Minimum Cost to Time Ratio with Application to a Ship Routing Problem," Technical Report No. 66-1, November 1966, Operations Research House, Stanford University.
- [2] Dantzig, G. B., "All Shortest Routes in a Graph," Technical Report No. 66-3, Operations Research House, Stanford University, November 1966.
- [3] Dantzig, G. B., "The Shortest Route Through a Network," Management Science, Vol. 6, No. 2, January, 1960, also in Linear Programming and Extensions, Princeton University Press, Princeton, N. J., 1963, pages 361-66.
- [4] Hu, T. C., "Revised Matrix Algorithms for Shortest Paths," IBM Watson Research Center, Research Paper, RC-1478, September 28, 1965.
- [5] Murchland, J. D., "Bibliography of the Shortest Route Problem," LSE-TWT-6, June 1966. (Revision)

FINDING A CYCLE IN A GRAPH WITH MINIMUM COST TO TIME RATIO  
WITH APPLICATION TO A SHIP ROUTING PROBLEM

BY

G. B. Dantzig\*, W. O. Blattner,\*\* and M.R. Rao\*\*

TECHNICAL REPORT NO. 66-1

November 1966

\*Operations Research House  
Stanford University  
Stanford, California

\*\*United States Steel Corporation

Research of G. B. Dantzig partially supported by Office of Naval Research,  
Contract ONR-N-00014-67-1-0112-0011, U. S. Atomic Energy Commission,  
Contract No. AT(04-3)-326 PA #18, and National Science Foundation Grant  
GP 6431; reproduction in whole or in part for any purpose of the United  
States Government is permitted.

## Finding a Cycle in a Graph with Minimum Cost to Time Ratio

### with Application to a Ship Routing Problem

by

G. B. Dantzig\*, W. O. Blattner\*\*, and M. R. Rao\*\*

Associated with each arc  $(i,j)$  of a graph are two numbers  $c_{ij}$  the "cost" and  $t_{ij}$  the "time" per unit flow. In our application the unit flow is a ship making one trip from  $i$  to  $j$  at a cost  $c_{ij}$  and taking  $t_{ij}$  hours. In another example, a vessel for hire can make a profit  $p_{ij}$  each time it goes from  $i$  to  $j$ ; eventually (if there are a finite number of ports) it must complete a cycle with a total profit  $P$  and a total lapsed time  $T$  where  $P$  is the sum of the profits and  $T$  is the sum of the times on the arcs of the cycle. For a maximum rate of return, the shipowners should use that cycle which maximizes the ratio of  $P/T$ . Later we shall describe a more complex linear programming model which we solve using a column generation scheme (a variant of the decomposition principle). The subproblem turns out to be one of finding a cycle in a graph that has the minimum ratio of total cost to total time.

Consider the following linear program:

Find  $\text{Min } z$ ,  $x_{ij} \geq 0$  such that

$$(1) \quad \sum_{i,j=1}^n c_{ij} x_{ij} = z$$

$$(2) \quad \sum_{i,j=1}^n t_{ij} x_{ij} = 1 \quad t_{ij} \geq 0$$

$$(3) \quad \sum_{i=1}^n x_{ij} - \sum_{k=1}^n x_{jk} = 0 \quad j = 1, 2, \dots, n$$

---

\*Stanford University

\*\*U. S. Steel

Theorem 1: Associated with an extreme minimizing solution to (1), (2), (3) is a cycle whose total cost to time ratio is minimum.

Proof: Let  $\bar{x}_{ij} = 1$  if  $(i,j)$  is an arc of some cycle and  $\bar{x}_{ij} = 0$  otherwise. Let  $\sum t_{ij} \bar{x}_{ij} = T$ , then  $x_{ij} = \bar{x}_{ij}/T$  satisfies (2) and (3) and  $z = C/T$  is the ratio of total costs,  $C = \sum \sum c_{ij} \bar{x}_{ij}$ , to total time  $T$ . Accordingly we can always associate with a cycle one of the solutions of (1), (2), (3).

Consider now the class of minimizing solutions to (1), (2), (3). We can now see that to an extreme minimizing solution corresponds a simple cycle. This follows because the flows  $x_{ij} \geq 0$  can be represented as a sum of simple circulations. If any of these circulations had by itself a lower ratio  $\sum c_{ij} x_{ij} / \sum t_{ij} x_{ij}$  than another one, the solution could not be optimal. Indeed an improved solution could be obtained by building up the circulation around that cycle with the lowest ratio and decreasing the flow around the one with a higher ratio. Nor could a solution be extreme if there were two simple cycles with the same ratio because one could represent such a solution as a convex combination of two others by first building up and then building down the circulation in one of the cycles while adjusting the other so (2) holds.

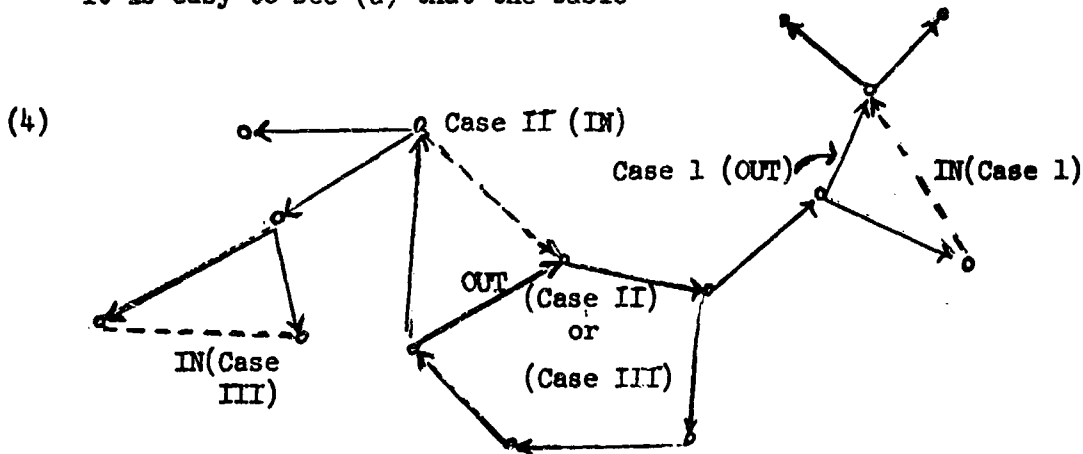
#### THE SIMPLEX ALGORITHM

A simple algorithm for solving (1), (2), (3) can be derived from the simplex method. A basis involves  $n$  columns (one equation is redundant). The corresponding arcs in the graph must consist of a tree and one out-of-tree arc. To see this we note that since a basis is non-singular, there must be at least one non-singular  $(n-1) \times (n-1)$  submatrix formed by deleting the row associated with the time equation and deleting some column of the basis. Non-singularity implies that the  $n-1$  arcs associated with the remaining columns form a tree. The arc associated with the variable of the deleted

column together with a subset of the arcs of the tree form a cycle.

In fact, this implies that every basic feasible solution to (2) and (3) must consist of one tree spanning all nodes and one simple cycle formed by a subset of the arcs of the tree augmented by one additional arc. This additional arc completes the cycle making possible the positive flow forced by the time equation.

It is easy to see (a) that the basic



variables other than the cycle variables have zero value in a basic solution, (b) that each node in the simple cycle has one cycle arc pointing into and the other away from it, and (c) the values of the cycle variables are the same and equal to  $1/T$  where  $T$  is the total time around the cycle.

It is also easy to compute the simplex multipliers (prices) associated with the basis. Indeed if we let  $\rho$  be the multiplier associated with the time equation (2) and let  $\Pi_j$  be those associated with the node equations (3), then for each arc  $(i,j)$  associated with a basic variable

$$\Pi_j - \Pi_i + \rho t_{ij} = c_{ij}$$

Summing these relations for all arcs  $(i,j) \in \text{Cycle}$  yields

$$(5) \quad \rho = C/T = \sum c_{ij} / \sum t_{ij} \quad (i,j) \in \text{cycle arcs}$$

Knowing  $\rho$ , one may arbitrarily choose the value of any one node (there is a redundant equation) and determine the remainder by

$$(6) \quad \pi_j = \pi_i + (c_{ij} - \rho t_{ij}) \quad (i,j) \in \text{tree arcs}$$

by branching out from the selected node along arcs  $(i,j)$  of the tree until all nodes are reached.

To obtain an improved solution the simplex multipliers are used to eliminate the basic variables from the cost equation. The resulting coefficients for the non-basic variables are

$$(7) \quad \delta_{ij} = (c_{ij} - \rho t_{ij}) - (\pi_j - \pi_i)$$

If all  $\delta_{ij} \geq 0$ , then the value of  $\rho$  given by (5) is the minimum cost to time ratio and the problem is solved.

If not, let

$$(8) \quad \delta_{pq} < 0 \quad \text{for some } (p,q)$$

We now make a special Inductive Assumption: at each iteration, there is a basic feasible solution consisting of a directed tree spanning out from a single node that is its root, augmented by one additional arc to form one simple cycle.<sup>†</sup>

---

<sup>†</sup> If a feasible cycle exists in the graph we can satisfy the inductive assumption by taking any node of the cycle as the root of the tree and spanning out from the nodes of the cycle using forward arcs to other nodes, and then iteratively, repeating the process with all nodes reached. If nodes still remain they can be reached by introducing high cost artificial arcs as required. If there is no feasible cycle in the graph there is obviously no feasible solution to the problem; if such is indeed the case this fact will be discovered by the algorithm that follows.

If  $(p,q)$  is the dotted arc marked "IN (Case I)" in the figure, then it is easy to see that entering  $(p,q)$  into the set of basic arcs does not form a new cycle and we must drop out of the basic set the one which is also directed into  $q$ . A new basic feasible set of arcs is obtained with the values of  $\Pi_i$  decreased to  $\Pi_i + \delta_{pq}$  for node  $i = q$  and all nodes  $i$  that are followers of  $q$  in the new tree. All other  $\Pi_i$  remain unchanged. In Cases II and III a new cycle is formed and we must drop out of the basis any one arc  $(r,s)$  which is in the old cycle and not in the new cycle.

Theorem 2: If the inductive assumption holds, and if  $(p,q)$  is entered into the basic set in place of the basic arc directed into  $q$ , then the inductive assumption holds after the change except when a new cycle is formed.

Theorem 3: If a new cycle is obtained as a result of changes in the basic set of arcs, its  $\rho^* = C/T$  ratio is less than the previous one.

Theorem 4: If  $d_{ij} = (c_{ij} - \rho t_{ij})$  is used as distances on arcs  $(i,j)$  in the graph, then any cycle in the graph whose sum of distances around the arcs of the cycle is negative has a lower  $C/T$  ratio.

The simplex method accordingly reduces down to finding a negative cycle in a graph when arc distances  $d_{ij}$  are given: Starting with  $\Pi_i = 0$  for some node of the cycle, the other  $\Pi_i$  are simply the distances from this origin node along arcs of the tree to node  $i$ . The simplex iterative process is seen to be the standard one for determining the shortest route from the origin to all others, terminating when either

$$(9) \quad d_{ij} - (\pi_j - \pi_i) \geq 0 \quad \text{for all } i, j$$

or on some iteration a negative cycle like in Cases II or II is found. In the former case, it is seen by summing (9) for all  $(i, j)$  around any given cycle that distance around any cycle is non-negative so that the  $\rho$  used to determine  $d_{ij} = c_{ij} - \rho t_{ij}$  is the best ratio. In the latter case, a new cycle with a lower value of  $\rho$  is obtained and new  $d_{ij}$  values are computed.

ALGORITHM: In the following algorithm the pairs  $(i, j)$  represent directed arcs defined in the graph:

0: Let  $S_0$  be any starting cycle. If none available, set

$$\rho_0 = \text{Max } [c_{ij}/t_{ij}] \text{ in Step 2 below}$$

1: For  $k = 0, 1, \dots$

2: Compute  $\rho_k = \sum c_{ij} / \sum t_{ij} \quad (i, j) \in S_k$

3: Compute  $d_{ij}^k = c_{ij} - \rho_k t_{ij} \quad \text{for all } i, j$

4: Set  $\pi_1^k = 0$  and set predecessor of node 1 as \*  
(meaning none).

Set  $\pi_i^k = \infty$  and set predecessor for  $i \neq 1$  as 1. ††

5: For each  $i = 1, 2, \dots, n$  form  $\delta_{ij}^k = d_{ij}^k + \pi_i^k - \pi_j^k$  for  
 $j = 1, 2, \dots, n \quad j \neq i$ .

(a) If  $\delta_{ij}^k \geq 0$  return to (5) and continue scanning  $j$  for fixed  $i$  and then repeat increasing  $i$  to  $i + 1$  until  $i = n, j = n - 1$ . If  $i = n, j = n - 1$  terminate.  
Cycle  $S_k$  is optimal.

---

††These are devices to initiate the computation without effort and to construct the starting directed tree necessary to satisfy the inductive assumption.

(b) If  $\delta_{ij}^k < 0$  go to (6).

6: Determine the nodes in reverse order along the route R from the origin to  $i$  by back tracing the predecessors of  $i$ .

(a) If  $j$  is not a predecessor of  $i$ , then change predecessor of  $j$  to  $i$  and replace value of  $\Pi_j^k$  by  $\Pi_j^k + \delta_{ij}^k$ , return to (5) with  $i = 1$ ,  $j = 2$ .

(b) If  $j$  is a predecessor of  $i$ , then let  $S_{k+1}$  be the cycle along the route R from node  $j$  to  $i$  and back to  $j$  along arc  $(i, j)$ . Return to 1 increasing  $k$  to  $k + 1$ .

For programming simplicity the above algorithm does not maintain a directed tree. If it is modified to do so, the nodes can be priced sequentially along the tree and the return from step 6 (a) to step (5) modified to take advantage of this.

# COMPUTATIONAL EXPERIENCE

<u>Set I</u>			
Problem	Nodes	Arcs	Seconds on IBM 1620 (including input-output)
A	4	12	3.60
B	4	12	2.16
C	5	20	3.96
D	4	7	2.52
E	6	13	6.84

<u>Set II</u>			
			(Excluding input-output)
F	5	20	1.80
G	10	90	7.92
H	15	210	14.04
I	20	380	33.84
J	25	600	36.00
K	30	870	103.68

For problems in Set II, the  $t_{ij}$  values for each arc were randomly generated integers between 10 and 60. Similarly, the  $c_{ij}$  values were randomly generated between 20 and 120.

We do not have an upper bound on the number of operations except the kind that one could derive from a standard proof of the simplex algorithm. In another paper where a variant of the scanning procedure given here is used an upper bound of  $(\text{Nodes} + 3)(\text{Arcs})$  additions-comparison operations is given for finding a negative cycle, [1].

### Application to a Ship Routing Problem

Amounts  $b_{ij}$  are required to be shipped from ports  $i$  to ports  $j$ . There are  $n$  ports (nodes). The shipping can either be done by charter at a cost  $v_{ij}$  per unit shipped or by using one of a fleet of  $m$  vessels under the control of the shipping company. If vessel  $k$  is used, the amount that it carries between  $(i,j)$  depends on the kind of ship and on the pattern of ports forming a cycle  $g$  that is assigned to the ship. We denote this by  $a_{kg}^{ij}$ . Thus if arc  $(i,j)$  is not part of cycle  $g$ , then  $a_{kg}^{ij} = 0$  and if it is its value is the capacity  $w_{ij}^k$  of the vessel. †††

Material balance equations: For  $i, j = 1, 2, \dots, n$

$$(10) \quad y_{ij} + \sum_{k=1}^n x_{kg} = 1 \quad \sum_g a_{kg}^{ij} x_{kg} = b_{ij}$$

where  $y_{ij}$  is amount chartered and  $x_{kg}$  is the number of times that ship  $k$  is employed in the  $g$ -th type cycle. We allow  $x_{kg}$  to have fractional values which we interpret as rate of use of the ship in some given period of time.

Vessel hour constraints:

$$(11) \quad \sum_g t_{kg} x_{kg} + s_k = h_k \quad k = 1, 2, \dots, m.$$

where  $h_k$  is the total hours available on the  $k$ -th vessel,  $s_k$  is the unused hours of the ship,  $t_{kg}$  is the time to complete one cycle of type  $g$ .

Objective to be minimized:

$$(12) \quad \sum_{(i,j)} v_{ij} y_{ij} - \sum_k c_k s_k = z$$

††† Dependence on  $(i,j)$  is possible if type of cargo on route  $(i,j)$  is different from that on other arcs. In case of airplanes, capacity depends on distance.

Here we are assuming that the cost to operate vessel is  $c_k$  per unit time used, hence there is a savings of  $c_k$  per hour not used.

In an ore shipment application which we were interested in there were too many possible cycles to explicitly list all the coefficients of the problem. Accordingly we decided to generate the column of coefficients as needed. Using  $y_{ij}$  and  $s_k$  as basic variables, one has a starting basic feasible solution. We now assume we have introduced into the basis several other columns and we have a set of simplex multipliers  $p_{ij}$  associated with (10) and  $q_k$  with (11). We wish to "price out" the column associated with  $x_{kg}$  and to find that column  $g$  for each  $k$  that prices out most negative. The relative cost coefficient of  $x_{kg}$  becomes

$$(13) \quad -q_k t_{kg} - \sum_{(i,j)} p_{ij} a_{kg}^{ij} = \sum_{(i,j) \in g} (-q_k t_{ij}^k - p_{ij} w_{ij}^k)$$

Our subproblem becomes one of choosing that cycle in the network of ship  $k$  for which (13) is a minimum. Since  $a_{kg}^{ij} = w_{ij}^k$  is the ship's capacity, if the arc  $(i,j)$  is used in the cycle  $g$  and zero otherwise, the sum in (13) is simply the sum of the ship capacities on arcs  $(i,j)$  weighted by  $p_{ij}$  and the times weighted by  $q_k$  around the cycle  $g$ . Note that  $t_{kg}$  is the sum of times on arcs  $(i,j)$  around the cycle. Unfortunately the problem in this form is that of finding a most negative cycle in a graph whose arc distances are given. This class of problems includes as a special case the difficult travelling salesman problem.

We got around this difficulty by a change of units. We set

$x_{kg} = \bar{x}_{kg}/t_{kg}$ . The relative cost coefficients for the new problem become

$$(14) \quad -q_k = (\sum_g p_{1j} w_{1j}^k / t_{kg})$$

where  $g$  denotes the  $(i,j) \in \text{cycle } g$

Since  $q_k$  for fixed  $k$  is constant and  $t_{kg} = \sum_g t_{1j}^k$  the subproblem becomes one of finding that cycle  $g^*$  that minimizes the ratio

$$(15) \quad (-\sum_g p_{1j} c_{1j}^k) / (\sum_g t_{1j}^k)$$

which fortunately, as we have seen, is a solveable problem!

#### REFERENCES

- [1] Dantzig, G. B., Blattner, W., and Rao, M. R., "All Shortest Routes from a Fixed Origin in a Graph," Research Report No. 66-2, Operations Research House, Stanford University, November 1966.

**SURVEY OF MATHEMATICAL PROGRAMMING**

by

**MICHEL L. BALINSKI**

at the

**American Mathematical Society Summer Seminar**

on the

**Mathematics of the Decision Sciences**

**Stanford University**

**July - August 1967**

# 1. Pivot Transformations

## The schema

$$(1.1) \quad \begin{array}{ccccccccc} & & & \dots & \eta^* & \dots & \eta & \dots & \\ \vdots & & & & \vdots & & \vdots & & \vdots \\ x^* & \dots & a & \dots & b & \dots & & & -y^* \\ \vdots & & \vdots & & \vdots & & & & \vdots \\ x & \dots & c & \dots & d & \dots & & & -y \\ \vdots & & \vdots & & \vdots & & & & \vdots \\ & & & \dots & -\xi^* & \dots & -\xi & \dots & \end{array}$$

conveniently exhibits two systems of linear equations, a

## row system

$$(1.1r) \quad \begin{array}{ccccccc} \vdots & & \vdots & & \vdots & & \vdots \\ \dots & + a\eta^* & + \dots & + b\eta & + \dots & = -y^* \\ \vdots & & \vdots & & \vdots & & \vdots \\ \dots & + c\eta^* & + \dots & + d\eta & + \dots & = -y \\ \vdots & & \vdots & & \vdots & & \vdots \end{array}$$

in which the dependent -y-labels (variables or numerical values) are expressed as linear combinations of the independent  $\eta$ -labels; and a

## column system

$$(1.1c) \quad \begin{array}{ccccccc} \vdots & & \vdots & & \vdots & & \vdots \\ \dots & + x^*a & + \dots & + xc & + \dots & = \xi^* \\ \vdots & & \vdots & & \vdots & & \vdots \\ \dots & + x^*b & + \dots & + xd & + \dots & = \xi \\ \vdots & & \vdots & & \vdots & & \vdots \end{array}$$

in which the dependent  $\xi$ -labels are expressed as linear combinations of the independent  $x$ -labels.

(1)

A pivot transformation with pivot entry  $a \neq 0$  simultaneously re-expresses the pair of linear systems by solving the  $-y^*$ -equation of the row system for  $\eta^*$  and the  $\xi^*$ -equation of the column system for  $x^*$ , and then using these equations to eliminate  $\eta^*$  and  $x^*$  as independent labels from the remaining row and column equations.  $y^*$  and  $\xi^*$  thereby become

independent labels. Solving for  $\eta^*$  in the row system

$$\dots + \eta^* + \dots + a^{-1}b\eta + \dots = -a^{-1}y^*$$

or

$$\dots a^{-1}y^* + \dots + a^{-1}b\eta + \dots = -\eta^*,$$

and hence, by substitution,

$$\dots c(\dots -a^{-1}y^* - \dots -a^{-1}b\eta - \dots) + \dots + d\eta + \dots = -y$$

or

$$\dots -ca^{-1}y^* + \dots + (d - ca^{-1}b)\eta + \dots = -y.$$

Solving for  $x^*$  in the column system

$$\dots + x^* + \dots + xca^{-1} + \dots = \xi^*a^{-1}$$

or

$$\dots + \xi^*a^{-1} - \dots - xca^{-1} - \dots = x^*$$

and hence, by substitution,

$$\dots + (\dots + \xi^*a^{-1} - \dots - xca^{-1} - \dots)b + \dots + xd = \xi$$

or

$$\dots + \xi^*a^{-1}b + \dots + x(d - ca^{-1}b) + \dots = \xi.$$

Therefore, the result of making a pivot transformation with pivot entry  $a \neq 0$  can be summarized in the transformation of the schema (1.1) into the schema (1.2)

$$(1.2) \quad \begin{array}{ccccccc} & & y^* & & \eta & & \dots \\ \vdots & & \vdots & & \vdots & & \vdots \\ \xi^* & \dots & a^{-1} & \dots & a^{-1}b & \dots & = -\eta^* \\ \vdots & & \vdots & & \vdots & & \vdots \\ x & \dots & -ca^{-1} & \dots & d-ca^{-1}b & \dots & = -y \\ \vdots & & \vdots & & \vdots & & \vdots \\ & \dots & -x^* & \dots & -\xi & \dots & \end{array}$$

In words the schema (1.2) is obtained from the schema (1.1) by

1. Exchanging labels corresponding to the pivot entry  $a$ ; namely, replacing  $\eta^*$  with  $y^*$ , and  $-y^*$  with  $-\eta^*$  in the row system, and replacing  $x^*$  with  $\xi^*$ , and  $\xi^*$  with  $x^*$  in the column system.

2. Keeping all remaining labels (e.g.,  $\eta$ ,  $-y$ ;  $x$ ,  $\xi$ ) unchanged.
3. Replacing the pivot  $a$  by its reciprocal  $1/a$ ; replacing each remaining entry in the pivot's row  $b$  by  $b/a$ ; replacing each remaining entry in the pivot's column  $c$  by  $-c/a$ ; and replacing each other entry  $d$  in  $b$ 's column and  $c$ 's row by  $(d - \frac{bc}{a})$ .

Clearly the only result of making a pivot transformation is to re-express or re-present a pair of linear systems in terms of different sets of independent and dependent labels.

Example. (\* denotes the pivot)

$$\begin{array}{c}
 \begin{array}{c} \eta_1 \quad \eta_2 \quad \eta_3 \quad \eta_4 \\ x_1 \quad \begin{array}{|ccc|} \hline 3 & -1 & -2 & 1 \\ \hline \end{array} \quad \begin{array}{l} = -y_1 \\ = -y_2 \\ = -y_3 \end{array} \\ x_2 \quad \begin{array}{|ccc|} \hline 2 & -2^* & 1 & 0 \\ \hline \end{array} \\ x_3 \quad \begin{array}{|ccc|} \hline 4 & 1 & 0 & -3 \\ \hline \end{array} \\ \hline \begin{array}{cccc} =t_1 & =t_2 & =t_3 & =t_4 \end{array}
 \end{array}
 \longrightarrow
 \begin{array}{c}
 \begin{array}{c} \eta_1 \quad y_2 \quad \eta_3 \quad \eta_4 \\ x_1 \quad \begin{array}{|ccc|} \hline 2 & -1/2 & -5/2 & 1 \\ \hline \end{array} \quad \begin{array}{l} = -y_1 \\ = -\eta_2 \\ = -y_3 \end{array} \\ t_2 \quad \begin{array}{|ccc|} \hline -1 & -1/2 & -1/2 & 0 \\ \hline \end{array} \\ x_3 \quad \begin{array}{|ccc|} \hline 5 & 1/2 & 1/2 & -3 \\ \hline \end{array} \\ \hline \begin{array}{cccc} =t_1 & -x_2 & =t_3 & =t_4 \end{array}
 \end{array}
 \end{array}$$

## 2. Dual Linear Programs.

The schema

$$\begin{array}{c}
 \begin{array}{c} (\geq 0) \qquad (\geq 0) \\ y_1 \quad \dots \quad y_N \quad 1 \\ \lambda_1 \quad \begin{array}{|ccc|} \hline a_{11} & \dots & a_{1N} \\ \hline \end{array} \quad \begin{array}{l} b_1 \\ \vdots \\ b_M \end{array} \\ \vdots \quad \begin{array}{|ccc|} \hline \vdots \\ \hline \end{array} \quad \begin{array}{l} \vdots \\ \vdots \end{array} \\ \lambda_M \quad \begin{array}{|ccc|} \hline a_{M1} & \dots & a_{MN} \\ \hline \end{array} \quad \begin{array}{l} b_M \\ \vdots \\ d \end{array} \\ 1 \quad \begin{array}{|ccc|} \hline c_1 & \dots & c_N \\ \hline \end{array} \quad \begin{array}{l} \\ \\ \\ d \end{array} \\ \hline \begin{array}{cccc} -x_1 & \dots & -x_N & = u \\ (\geq 0) & & (\geq 0) & (\max) \end{array}
 \end{array}
 \end{array}
 \begin{array}{l} = 0 \\ \\ \\ = 0 \\ \\ = v \text{ (min)} \end{array}$$

conveniently exhibits a dual pair of linear programs in standard form

Row (or Primal) Program

Minimize

$$v = c_1 y_1 + \dots + c_N y_N + d$$

constrained by

$$a_{11} y_1 + \dots + a_{1N} y_N + b_1 = 0$$

(2.lx,c)

$$a_{M1} y_1 + \dots + a_{MN} y_N + b_M = 0$$

$$y_1 \geq 0, \dots, y_N \geq 0$$

Column (or Dual) Program

Maximize

$$u = \lambda_1 b_1 + \dots + \lambda_M b_M + d$$

constrained by

$$\lambda_1 a_{11} + \dots + \lambda_M a_{M1} + c_1 = x_1 \geq 0$$

$$\lambda_1 a_{1N} + \dots + \lambda_M a_{MN} + c_N = x_N \geq 0$$

**Theorem 1.** If  $(v; y_1, \dots, y_N)$  satisfies the row equations and  $(u; \lambda_1, \dots, \lambda_M; x_1, \dots, x_N)$  satisfies the column equations then

$$(2.2) \quad x_1 y_1 + \dots + x_N y_N = v - u.$$

**Theorem 2.** If  $(v^0; y_1^0, \dots, y_N^0)$  satisfies the row constraints (2.1r) (is "row feasible") and  $(u^1; \lambda_1^1, \dots, \lambda_M^1; x_1^1, \dots, x_N^1)$  satisfies the column constraints (2.1c) (is "column feasible") then

$$u^0 \leq \text{minimum } v \leq v^1$$

and

$$u^0 \leq \text{maximum } u \leq v^1.$$

**Corollary 1.** If feasible  $(v^0; y_1^0, \dots, y_N^0)$  and  $(u^1; \lambda_1^1, \dots, \lambda_M^1; x_1^1, \dots, x_N^1)$  can be found such that  $v^0 = u^1$ , then they constitute (optimal) solutions to row and column programs. This can happen only if (by (2.2))  $x_1 y_1 = 0$  for all  $i$ , that is, only if  $x_i = 0$  and/or  $y_i = 0$ .

### 3. Reduction to Canonical Form.

In order to solve a dual pair of linear programs (2.1) which are in standard form the first task is to eliminate the unrestricted  $\lambda$ -variables from the column program, and re-express the row program in terms of the smallest possible set of independent  $y$ -variables. This is accomplished by using the

**Reduction Rule.** Make pivot transformations on (2.1) and its successive representations with pivot entries corresponding to 0-dependent and  $y$ -independent row program labels and  $x$ -dependent and  $\lambda$ -independent column program labels until no longer possible.

When such a pivot entry cannot be chosen then a representation has been obtained whose schema is such that either

- (a) every entry which corresponds to 0-dependent and  $y$ -independent row program labels and to  $x$ -dependent and  $\lambda$ -independent column program labels is a zero and thus cannot be chosen as a pivot; or
- (b) no row program dependent labels are 0's and no column program independent labels are  $\lambda$ 's.

If (a) occurs we must have obtained a schema of the form

$$\begin{array}{rcccl}
 & y'_{m+1} & \dots & y'_{m+n} & 0 & \dots & 0 & 1 & \\
 x'_1 & a'_{11} & \dots & a'_{1n} & a'_{1n+1} & \dots & a'_{1N} & b'_1 & = -y'_1 \\
 \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots & \vdots \\
 x'_m & a'_{m1} & \dots & a'_{mn} & a'_{mn+1} & \dots & a'_{mN} & b'_m & = -y'_m \\
 \hline
 \lambda'_{m+1} & 0 & \dots & 0 & a'_{m+1,n+1} & \dots & a'_{m+1,N} & b'_{m+1} & = 0 \\
 \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots & \vdots \\
 \lambda'_M & 0 & \dots & 0 & a'_{Mn+1} & \dots & a'_{MN} & b'_M & = 0 \\
 \hline
 1 & c'_1 & \dots & c'_n & c'_{n+1} & \dots & c'_N & d' & = v \\
 & = x'_{m+1} & & = x'_{m+n} & = \lambda'_1 & & = \lambda'_m & = u & 
 \end{array}
 \quad (3.1)$$

where the primed variables are a rearrangement of the original variables in (2.1) and the primed entries are determined by the succession of preceding pivot steps. If any of the entries  $b'_{m+1}, \dots, b'_M$  is different from zero then the row program constraints are incompatible, for some row equation corresponding to a dependent 0-label would read "a non-zero quantity equals zero." Then no optimal solutions exist. Otherwise, if all  $b'_{m+1}, \dots, b'_M = 0$ , the row equations corresponding to dependent 0-labels read "zero equals zero" and can thus be omitted. Clearly the columns corresponding to the independent 0-labels can also be omitted from the point of view of the row program. These same rows and columns may be omitted from the point of view of the column program: the equations corresponding to the unrestricted  $\lambda'_1, \dots, \lambda'_m$  can be put aside since they represent no constraints; then, the coefficients corresponding to the independent  $\lambda'_{m+1}, \dots, \lambda'_M$  are all zeros and hence their rows may be omitted. Notice that the argument made from the column program point of view holds whether or not all  $b'_{m+1}, \dots, b'_M$  are zero; however, if some of these are not zero then, clearly, even if there exists a column program feasible solution, the value of the objective, maximise  $u$ , can be made arbitrarily large (why?).

In summary if (a) occurs and all  $b'_{m+1}, \dots, b'_M$  are zero we obtain a smaller representation for the dual pair of linear programs (2.1) whose schema is

(3.2)

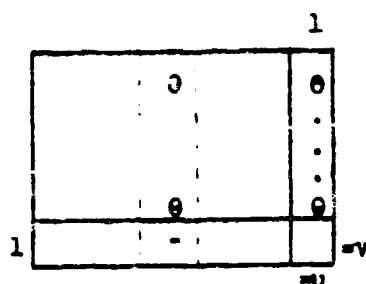
Alternative (b) is a special case of (a) with  $m = M$ , i.e., all  $\lambda$ -labels are dependent, all 0-labels are independent. Therefore, for either alternative, we can state

#### 4. The Main Theorem of Linear Programming

(4.2)

exhibiting optimal solutions  
to both programs;

(4.b)



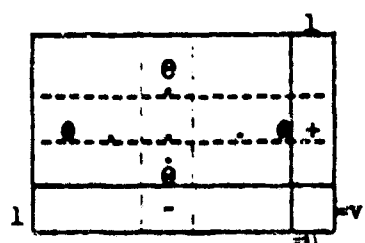
exhibiting a feasible solution to the row program and the unboundedness (from below) of  $v$ ; and showing the noncompatibility of the column constraints;

(4.c)



exhibiting a feasible solution to the column program and the unboundedness (from above) of  $u$ ; and showing the noncompatibility of the row constraints;

(4.d)



showing the noncompatibility of both row and column constraints.

(Proof of Theorem 4 will be given below)

(2)

**Corollary 2.** If there exist feasible solutions to both row and column programs then there must exist optimal solutions.

**Proof:** If there exist feasible solutions to both programs then cases (b), (c), (1) cannot occur, and hence case (a) must occur.

Corollary 2 is usually referred to as the fundamental duality theorem of linear programming.

### 5. Simplex Methods (3)

A simplex method for solving a pair of dual linear programs (3.2) in canonical form is a finite sequence of schemata or equivalent representations for the programs obtained by successive pivot steps, with prescribed pivot entry choice rules, which obtain a schema exhibiting optimal solutions to both programs, or the non-compatibility of the row and/or the column constraints.

(4)

A row (or primal) simplex method is a simplex method beginning with a schema exhibiting row feasibility with pivot steps which maintain row feasibility in each succeeding schema. Thus, at any stage, a schema

$$(5.1) \quad \begin{array}{c|ccc|c} & y_{m+1}' & \dots & y_{m+n}' & \\ \hline x_1' & a_{11}' & \dots & a_{1n}' & b_1' = -y_1' \\ \vdots & \vdots & & \vdots & \vdots \\ x_m' & a_{m1}' & \dots & a_{mn}' & b_m' = -y_m' \\ \hline 1 & c_1' & \dots & c_n' & d' = v \\ \hline & -x_{m+1}' & \dots & -x_{m+n}' & = u \end{array}$$

with  $b_1' \leq 0, \dots, b_m' \leq 0$  is at hand.

A row (or primal) pivot choice rule is as follows:

If a schema (5.1) does not exhibit optimal solutions to both programs (form 4.a) there must exist a  $c_j' < 0$  for some  $j$ . Either (i) every entry in the column of  $c_j' < 0$  is nonpositive or (ii) there exist positive entries.

(i) The schema is in form (4.b)

(ii) Choose as pivot entry  $a_{kj}' > 0$  satisfying

$$\frac{b_k'}{a_{kj}'} = \max_{s' a_{sj}' > 0} \frac{b_s'}{a_{sj}'}$$

Notice that if  $b_k' < 0$  a new row feasible solution is exhibited after pivoting which gives a value to  $v$  strictly less than the previous exhibited value of  $v$ .

Example. Solve the linear programs

$$\begin{array}{c|ccc|c} & y_3 & y_4 & y_5 & 1 \\ \hline x_1 & 3 & 4 & 1 & -3 = -y_1 \\ x_2 & 1 & 3^* & 2 & -1 = -y_2 \\ \hline 1 & -3 & -6 & -2 & 0 = v \text{ (min)} \\ \hline & -x_3 & -x_4 & -x_5 & = u \end{array} \quad \begin{array}{l} (\geq 0) \\ (\leq 0) \\ (\geq 0) \end{array}$$

$$\begin{array}{c}
 \begin{array}{ccccc}
 & y_3 & y_2 & y_5 & 1 \\
 x_1 & 5/3 & -4/3 & -5/3 & -5/3 & = -y_1 \\
 x_4 & 1/3^* & 1/3 & 2/3 & -1/3 & = -y_4 \\
 1 & -1 & 2 & 2 & -2 & = v \\
 & =x_3 & =x_2 & =x_5 & =u
 \end{array}
 \longrightarrow
 \begin{array}{ccccc}
 & y_4 & y_2 & y_5 & 1 \\
 x_1 & -5 & -3 & -5 & 0 & = -y_1 \\
 x_3 & 3 & 1 & 2 & -1 & = -y_3 \\
 1 & 3 & 3 & 4 & -3 & = v \\
 & =x_4 & =x_2 & =x_5 & =u
 \end{array}
 \end{array}$$

(5)

A column (or dual) simplex method is a simplex method beginning with a schema exhibiting column feasibility with pivot steps which maintain column feasibility in each succeeding schema. Thus, at any stage, a schema (5.1) with  $c'_1 \geq 0, \dots, c'_n \geq 0$ , is at hand.

A column (or dual) pivot choice rule is as follows:

If a schema (5.1) does not exhibit optimal solutions to both programs (form 4.a) there must exist a  $b'_i > 0$  for some  $i$ . Either (i) every entry in the row of  $b'_i > 0$  is nonnegative or (ii) there exist some negative entries.

(i) The schema is in form (4.c).

(ii) Choose as pivot entry  $a'_{ij} < 0$  satisfying

$$\frac{c'_i}{a'_{ij}} = \max_{a'_{is} < 0} \frac{c'_s}{a'_{is}}$$

Notice that if  $c'_i > 0$  a new column feasible solution is exhibited after pivoting which gives a value to  $u$  strictly greater than the previous exhibited value of  $u$ .

Example. Solve the linear programs

$$\begin{array}{c}
 (\geq 0) \\
 \begin{array}{ccccc}
 & y_1 & y_2 & y_3 & 1 \\
 x_4 & -3 & -1 & 0 & 3 & = -y_4 \\
 x_5 & -4 & -3^* & -1 & 6 & = -y_5 \quad (\leq 0) \\
 x_6 & -1 & -2 & +1 & 2 & = -y_6 \\
 1 & 2 & 1 & 1 & 0 & = v \text{ (min)} \\
 & =x_1 & =x_2 & =x_3 & =u \\
 & & (\geq 0) & (\text{max})
 \end{array}
 \end{array}$$

$$\begin{array}{ccccc}
 & y_1 & y_2 & y_3 & 1 \\
 x_4 & -5/3^* & -1/3 & 1/3 & 1 & = -y_4 \\
 x_2 & 4/3 & -1/3 & 1/3 & -2 & = -y_2 \\
 x_6 & 5/3 & -2/3 & 5/3 & -2 & = -y_6 \\
 1 & 2/3 & 1/3 & 2/3 & 2 & = v \\
 & =x_1 & =x_5 & =x_3 & =u
 \end{array}$$

$$\begin{array}{ccccc}
 & y_4 & y_5 & y_3 & 1 \\
 x_1 & -3/5 & 1/5 & -1/5 & -3/5 & = -y_1 \\
 x_2 & 4/5 & -3/5 & 3/5 & -6/5 & = -y_2 \\
 x_6 & 1 & -1 & 2 & -1 & = -y_6 \\
 1 & 2/5 & 1/5 & 4/5 & 12/5 & = v \\
 & =x_4 & =x_5 & =x_3 & =u
 \end{array}$$

The row and column simplex methods described above can be applied only to a pair of programs whose schema exhibits, respectively, a row or a column feasible solution. This is not always the case, of course. A method for obtaining a schema exhibiting a row feasible solution (if one exists) can be described as follows. (6)

Suppose that the schema (5.1) does not exhibit a row feasible solution, i.e.,  $b'_i > 0$  for some  $i$ . Assume that the nonpositive  $b'_i$  are  $b'_1 \leq 0, \dots, b'_k \leq 0$  (this is not a limiting assumption for the rows of (5.1) could be rearranged to accomplish this). Then, by replacing the  $b'_i$  with their signs, the schema (5.1) can be exhibited as

$$(5.2) \quad \begin{array}{c} \begin{array}{ccccccc} & y'_{m+1} & \cdot & \cdot & \cdot & y'_{m+n} & 1 \\ x'_1 & a'_{11} & \cdot & \cdot & \cdot & a'_{1n} & \theta & = -y'_1 \\ \vdots & & & & & & & \vdots \\ x'_k & & & & & & \theta & = -y'_k \\ x'_{k+1} & & & & & & + & = -y'_{k+1} \\ \hline \vdots & & & & & & & \vdots \\ x'_m & a'_{m1} & \cdot & \cdot & \cdot & a'_{mn} & + & = -y'_m \\ 1 & c'_1 & \cdot & \cdot & \cdot & c'_n & d' & = v \\ & -x'_{m+1} & \cdot & \cdot & \cdot & -x'_{m+n} & = u \end{array} \end{array}$$

In (5.2) the sub-schema above the double line can be thought of as specifying a pair of subprograms in canonical form in which a row feasible solution obtains. Therefore, a row simplex method pivot choice rule can be used directly in (temporarily) minimizing  $-y'_{k+1}$  subject to the constraints as specified above the  $-y'_{k+1}$  row. If, after one or several pivot steps, the sign of the entry corresponding to  $b'_{k+1}$  becomes nonpositive, then one more row with "correct" sign has been generated and there is a larger subschema (perhaps the entire schema) which can be thought of as specifying a pair of subprograms in which a row feasible solution obtains. Otherwise, one of the two following forms can be reached (by Theorem 4):

(5.3a)

				1	
	0			0	
	0			0	
	0			0	
	-*			+	$= -y_{k+1}^i$
1					$= v$
					$= u$

(5.3b)

				1	
	0			0	
	0			0	
	0			0	
	0	0	0	0	$= -y_{k+1}^i$
1					$= v$
					$= u$

If (5.3a) occurs a pivot step with pivot entry starred leads to a larger subschema in which a row feasible solution obtains (why?). If (5.3b) occurs no row feasible solution to the (complete) row constraints exists.

Example. Obtain a schema exhibiting a row feasible solution to

		$(\geq 0)$			
		$y_1$	$y_2$	$y_3$	
$x_4$	-3	-1*	0	1	$= -y_4$
$x_5$	-4	-3	-1	6	$= -y_5$
$x_6$	-1	-2	1	3	$= -y_6$
1	-1	1	1	0	$= v$ (min)
	$= x_1$	$= x_2$	$= x_3$	$= u$	
		$(\geq 0)$	$(\max)$		

	$y_1$	$y_4$	$y_3$	1	
$x_2$	3	-1	0	-1	$= -y_2$
$x_5$	5	-3	-1*	3	$= -y_5$
$x_6$	5	-2	1	1	$= -y_6$
1	-4	1	1	1	$= v$
	$= x_1$	$= x_4$	$= x_3$	$= u$	

	$y_1$	$y_4$	$y_5$	1	
$x_2$	3	-1	0	-1	$= -y_2$
$x_3$	-5	3*	-1	-3	$= -y_3$
$x_6$	10	-5	1	4	$= -y_6$
1	1	-2	1	4	$= v$
	$= x_1$	$= x_4$	$= x_5$	$= u$	

	$y_1$	$y_3$	$y_5$	1	
$x_2$	4/3	1/3	-1/3	-2	$= -y_2$
$x_4$	-5/3	1/3	-1/3	-1	$= -y_4$
$x_6$	5/3	5/3	-2/3	-1	$= -y_6$
1	-7/3	2/3	1/3	2	$= v$
	$= x_1$	$= x_3$	$= x_5$	$= u$	

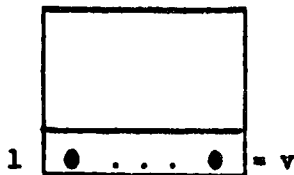
A method for obtaining a column feasible solution (if one exists) can be described analogously (the description is left to the reader).

#### 6. Proof of the Main Theorem.<sup>(7)</sup>

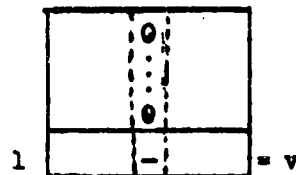
Suppose, as in the statement of the theorem, that a dual pair of linear programs in canonical form (3.2) is given with  $m + n + 2$  the number of rows plus columns. Before proceeding to a proof notice that two corollaries are immediate consequences of the theorem.

Corollary 3. Given a dual pair of linear programs in canonical form (3.2) there exists a finite succession of pivot transformations which obtain a representation for the programs whose schema has exactly one of the following two forms (where the identity of the column with label 1 is not distinguished):

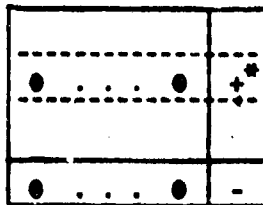
(6.e)



(6.f)



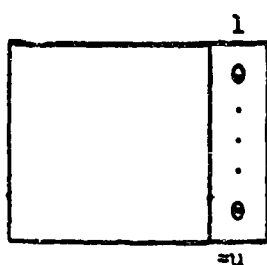
Proof. If (4.a), (4.b) or (4.d) hold then either (6.e) or (6.f) obtains. If (4.c) holds, but (6.e) does not obtain then



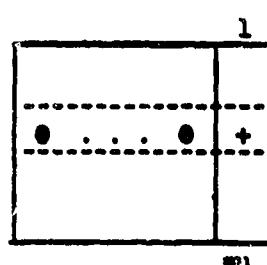
obtains; but with one pivot step with pivot entry designated, (6.4) obtains.

Corollary 4. Given a dual pair of linear programs in canonical form (3.2) there exists a finite succession of pivot transformations which obtain a representation for the programs whose schema has exactly one of the following two forms (where the identity of the row with label 1 is not distinguished):

(6.g)



(6.h)



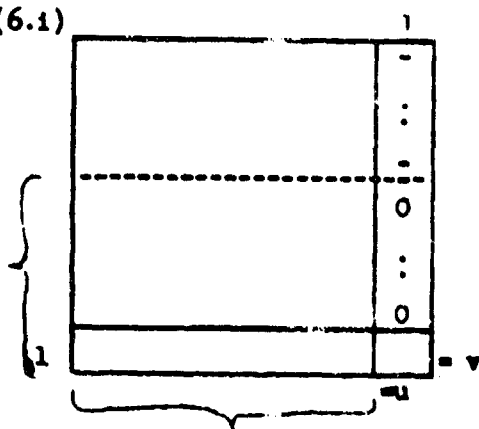
or

Proof. The proof is analogous to that for corollary 3 and is left to the reader.

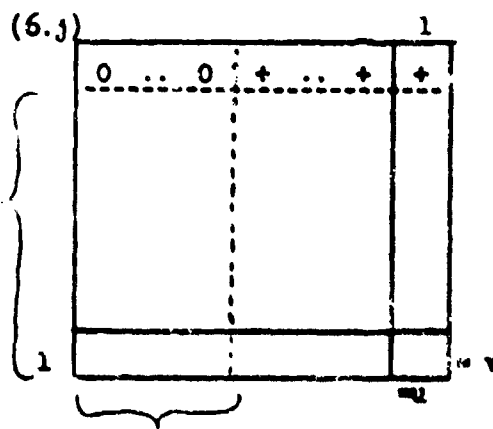
The proof of the Main Theorem is accomplished by induction on the number of rows plus columns. If there is either only one row or only one column in (3.2), one of the forms of the theorem obtains. Suppose, then, the theorem true, and hence Corollaries 3 and 4 true, for the number of rows plus columns less than  $m + n + 2$ .

Applying the inductive hypothesis to (3.2) with the last row ignored we must obtain, by corollary 4, either (after rearrangement of rows and columns according to signs)

(6.i)



(6.j)





;



This completes the proof. Notice, however, that to every application of the inductive hypothesis there corresponds a constructive computational set of rules for choice of pivot entry which achieve the same results. These rules are row and column simplex method pivot choice rules applied to appropriate subsets of rows and columns. (8)

## 7. Matrix Games

A matrix game  $A$  (or two-person zero-sum game in normalized form) is specified by any matrix of real numbers

$$(7.1) \quad A = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mn} \end{bmatrix}$$

with the rule that in a play of the game players I and II simultaneously choose some row  $i$  ( $i = 1, \dots, m$ ) and column  $j$  ( $j = 1, \dots, n$ ), respectively, with the result that II pays I an amount  $a_{ij}$ .

A pure strategy for player I is the choice of some one row of  $A$ ; a pure strategy for player II is the choice of some one column of  $A$ . A mixed strategy for player I is a probability vector  $X = (x_1, \dots, x_m)$ ,  $x_i \geq 0$  all  $i$ ,  $\sum x_i = 1$  (in a great many plays of the game I chooses row  $i$  with probability  $x_i$ ); a mixed strategy for player II is a probability vector  $Y = (y_1, \dots, y_n)$ ,  $y_j \geq 0$  all  $j$ ,  $\sum y_j = 1$  (in a great many plays of the game II chooses column  $j$  with probability  $y_j$ ).

If  $X = (x_1, \dots, x_m)$  and  $Y = (y_1, \dots, y_n)$  are any two mixed strategies for players I and II, respectively, the schema

$$(7.2) \quad \begin{array}{ccccc} & y_1 & & y_n & \\ x_1 & \boxed{a_{11} \quad \dots \quad a_{1n}} & = & s_1 & \\ \vdots & \vdots & & \vdots & \\ x_m & \boxed{a_{m1} \quad \dots \quad a_{mn}} & = & s_m & \\ & y_1 & \dots & y_n & \end{array}$$

conveniently exhibits, in the column system, players I's expected gains

$$\begin{array}{rcl}
 & x_1 a_{11} + \dots + x_m a_{m1} = g_1 \\
 (7.2, I) & \vdots & \vdots \\
 & x_1 a_{1n} + \dots + x_m a_{mn} = g_n
 \end{array}$$

against each of II's columns (or possible pure strategies); and, in the row system, player II's expected losses

$$\begin{array}{rcl}
 & a_{11} y_1 + \dots + a_{1n} y_n = l_1 \\
 (7.2, II) & \vdots & \vdots \\
 & a_{m1} y_1 + \dots + a_{mn} y_n = l_m
 \end{array}$$

against each of I's rows (or possible pure strategies).

The rationale of the theory of games requires that player I choose his mixed strategy  $X$  so as to maximize his minimum expected gain against any choice of column by II, and that player II choose his mixed strategy  $Y$  so as to minimize his maximum expected loss against any choice of row by I. Letting

$$(7.3) \quad u = \min_j g_j \quad \text{and} \quad v = \max_i l_i ;$$

this means player I's objective is to choose an  $X$  to maximize  $u$ , and player II's objective is to choose a  $Y$  to minimize  $v$ . Clearly  $u \leq v$ . Thus if an  $X$  and a  $Y$  can be found such that  $u = v$  they must constitute optimal strategies. The common value  $u = v$  is then called the value of the game  $A$ .

The objectives of the players may be formulated as linear



Two successive pivot transformations on the starred +1 and -1 of (7.4) obtain the representation for the programs exhibited in the schema

where

$$a_{ij}' = a_{ij} - a_{in} \quad a_{nj} + a_{nn} \quad (i \neq n, j \neq n)$$

(7.6) and

$$a_{in}' = a_{in} - a_{nn} \quad (i \neq n), \quad a_{nj}' = a_{nj} - a_{nn} \quad (j \neq n).$$

But (7.5) exhibits equivalent representations of the linear programs (7.4,I) and (7.4,II) as a dual pair of linear programs in canonical form. <sup>(?)</sup> This observation permits us to state

Theorem 5<sup>(10)</sup> (Minimax Theorem). There always exist optimal mixed strategies  $X$  and  $Y$  for players I and II, respectively, for which  $u = v$ .

Proof. The dual pair of linear programs (7.4,I) and (7.4,II) each have feasible solution, as is easily verified directly. Therefore, by Theorem 4, form (4.a) must obtain, establishing the theorem.

To find optimal mixed strategies and the value of the game A we need only to solve the dual pair of linear programs exhibited in (7.4) or (7.5) by applying a simplex method. Notice, however, that if  $a_{mn} = \min_j a_{mj}$  then (7.5) exhibits a column feasible solution, while if  $a_{mn} = \max_i a_{in}$  then (7.5) exhibits a row feasible solution. Therefore, if the rows or columns of A are rearranged so that one of these conditions hold a column or a row simplex method can be applied directly.

Furthermore, if

$$(7.7) \quad \min_j a_{mj} = a_{mn} = \max_i a_{in}$$

then optimal solutions obtain in (7.5) with  $x_m = 1$ ,  $y_n = 1$  and  $u = v = a_{mn}$  that is, the optimal strategies are pure strategies. The entry  $a_{mn}$  is called a saddle point, and we can state

Theorem 6. If  $a_{kl}$  is a saddle point of A, i.e., if

$$(7.8) \quad a_{il} \leq a_{kl} \leq a_{kj} \quad (\text{all } i, j)$$

then optimal pure strategies exist with  $x_k = 1$  and  $y_l = 1$ , and the value of the game is  $a_{kl}$ .

Examples.

Solve the matrix game A specified by

$$A = \begin{bmatrix} 0 & -1 & 1 & -1 & 7 \\ -7 & 1 & -3 & 2 & -8 \\ 3 & \textcircled{2} & 2 & 6 & 3 \\ 2 & \textcircled{2} & 3 & 4 & 5 \end{bmatrix}$$



	$t_2$	$t_3$	$y_2$	1	
$s_3$	1/2	-1/2	3	-1/2	$= -y_3$
$x_1$	5	-6	28*	0	$= -t_1$
$s_1$	-1/2	1/2	-2	-1/2	$= -y_1$
1	-1	2	-7	0	$= v$
	$=x_2$	$=x_3$	$=s_2$	$=u$	

	$t_2$	$t_3$	$t_1$	1	
$s_3$	-1/28	4/28	-3/28	-1/2	$= -y_3$
$s_2$	5/28	-6/28*	1/28	0	$= -y_2$
$s_1$	-4/28	2/28	2/28	-1/2	$= -y_1$
1	1/4	1/2	1/4	0	$= v$
	$=x_2$	$=x_3$	$=x_1$	$=u$	

Thus optimal mixed strategies are

$$X = (1/4, 1/4, 1/2, 0) \quad \text{and} \quad Y = (1/2, 0, 1/2, 0, 0)$$

and the value of the game is zero (it is fair). Notice, however, that another pair of optimal mixed strategies can be found by pivoting as indicated to obtain

	$t_2$	$y_2$	$t_1$	1	
$s_3$	1/2	4/6	-1/12	-1/2	$= -y_2$
$x_3$	-5/6	-28/6	-1/6	0	$= -t_3$
$s_1$	-1/2	2/6	1/12	-1/2	$= -y_1$
1	2/3	14/6	1/3	0	$= v$
	$=x_2$	$=s_2$	$=x_1$	$=u$	

Thus optimal mixed strategies are

$$X = (1/3, 2/3, 0, 0) \quad \text{and} \quad Y = (1/2, 0, 1/2, 0, 0).$$

Therefore

$$X = \alpha(1/3, 2/3, 0, 0) + (1-\alpha)(1/4, 1/4, 1/2, 0), \quad 0 \leq \alpha \leq 1$$

constitutes an optimal mixed strategy for player I for any choice of  $\alpha$  between zero and one. (Why?).

Footnotes.

- (1) Also known as complete elimination, or Gauss-Jordan elimination. See: E. Stiefel, "Note on Jordan Elimination, Linear Programming, and Tehebycleff Approximation", Numerische Mathematik vol. 2 (1960), pp. 1-17.
- (2) Corollary 2 is known as the "fundamental duality of linear programming". Its first explicit statement is contained in D. Gale, H. W. Kuhn, and A.W. Tucker, "Linear Programming and the Theory of Games", in Activity Analysis of Production and Allocation, (edited by T. C. Koopmans), John Wiley and Sons, Inc., New York, 1951. The notion of a duality theory arose from the Minimax Theorem of J. von Neumann (see footnote (10)).
- (3),(4) The discovery of a (primal) simplex method (1947) is due to George B. Dantzig. His original paper "Maximization of a Linear Function of Variables Subject to Linear Inequalities" is contained in Activity Analysis of Production and Allocation.
- (5) A dual simplex method was first explicitly advanced by C. C. Lemke in "The Dual Method of Solving the Linear Programming Problem", Naval Research Logistics Quarterly, vol. 1, (1954), pp. 36-37.
- (6) This method is described by M. L. Balinski and R. E. Gomory in "A Mutual Primal-Dual Simplex Method" to appear in Proceedings of Symposium on Mathematical Programming held in Chicago, June, 1962.
- (7),(8) The proof given here depends on the basic idea advanced in the paper referred to in (6), where the proof is completely constructive. The inductive proof found here is due to A. W. Tucker.
- (9) The formulation of the players' problems as linear programs given here follows A. W. Tucker's, "Solving a Matrix Game by Linear Programming", IBM Journal of Research and Development, vol. 4 (1960), pp. 507-517.
- (10) The Minimax Theorem was discovered and first proved by J. von Neumann in "Zur Theorie der Gesellschaftsspiele", Mathematische Annalen, vol. 100 (1928), pp. 295-320.

# ERRATA

for

## Pivot Transformations, Dual Linear Programs, and Simplex Methods

	Instead of:	Read:
Page 4. Lines 11, 12.	$x_i y_i = 0$ $x_i = 0$ $y_i = 0$	$x_i^1 y_i^0 = 0$ $x_i^1 = 0$ $y_i^0 = 0$
Page 12. Line 5.	to	of
Line 2 from bottom.	(6.4)	(6.e)
Page 14. Line 4.	(5.3a, b)	(5.3a, b)
Line 9.	pf	of
Page 15. Line 7 from bottom.	row	column
Bottom line.	appropriate	appropriate
Page 19. Under schema (7.5) insert		$(\geq 0)$
Line 9 from bottom.	$(i \neq m, j \neq n)$	$(i \neq m, j \neq n)$
Page 20. Line 15.	$a_{mn}$ that is,	$a_{mn}$ that is,
Line 19.	exist	exist
Page 21. Line 7.	more by	more or at least not less by
Footnotes. Line 3.	Tchebycleff	Tchebycheff
Line 18.	C. L. Lemke	C. E. Lemke

**NONLINEAR PROGRAMMING**

by

**TERRY ROCKAFELLAR**

at the

**American Mathematical Society Summer Seminar**

on the

**Mathematics of the Decision Sciences**

**Stanford University**

**July - August 1967**

## DUALITY IN NONLINEAR PROGRAMMING

R.T. Rockafellar\*

### 1. Introduction

A nonlinear program in  $n$  variables is usually described as a problem of minimizing (or maximizing) a quantity  $f_0(x)$  subject to constraints  $f_1(x) \leq 0, \dots, f_m(x) \leq 0$ , where  $f_0, \dots, f_m$  are certain real-valued functions of the vector  $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ . The problem may be interpreted broadly or narrowly, however.

In the narrower sense, one is only interested in the infimum of a certain function given on a subset  $S$  of  $\mathbb{R}^n$ . The elements  $x$  of the subset  $S$  are the so-called feasible solutions to the problem. Typical questions are the following. Is the infimum finite? Do there exist optimal solutions, i.e. feasible solutions at which the infimum is attained? Is there only one optimal solution? One seeks conditions which guarantee "yes" answers to these questions, as well as algorithms for actually computing the infimum and optimal solutions.

In the broader sense of the problem, one is also concerned with the sensitivity of the infimum and optimal solutions to slight changes in the constraints. This is where duality and Lagrange multipliers come in. Let  $p(u_1, \dots, u_m)$  denote the infimum of  $f_0(x)$  subject to  $f_1(x) \leq u_1, \dots, f_m(x) \leq u_m$ .

---

\*This work was supported in part by grant AF-AFOSR-1202-67 from the Air Force Office of Scientific Research.

One is interested in the properties of  $p$  as a function of the perturbation  $u = (u_1, \dots, u_m) \in \mathbb{R}^m$  near  $u = 0$ . For instance, is  $p$  continuous or differentiable at  $u = 0$ ?

It is especially important to look for numbers  $u_1^*, \dots, u_m^*$  such that

$$(1.1) \quad p(u_1, \dots, u_m) \geq p(0, \dots, 0) - u_1^* u_1 - \dots - u_m^* u_m, \quad \forall (u_1, \dots, u_m) \in \mathbb{R}^m.$$

Such numbers can be interpreted as "equilibrium prices," if the objective function  $f_0$  is interpreted as a cost function. Suppose that in trying to minimize cost we are allowed to perturb the given problem by any amount  $(u_1, \dots, u_m)$ , but that this perturbation must be paid for, the price being  $u_i^*$  per unit of variable  $u_i$ . The minimal cost attainable in the problem perturbed by  $(u_1, \dots, u_m)$ , plus the cost of this perturbation, is

$$p(u_1, \dots, u_m) + u_1^* u_1 + \dots + u_m^* u_m.$$

If the prices satisfy (1.1), this is never less than the minimal cost  $p(0, \dots, 0)$  in the unperturbed problem, so all the incentive for perturbation is neutralized and there is an "equilibrium".

Observe that (1.1) is satisfied if and only if

$$f_0(x) + u_1^* u_1 + \dots + u_m^* u_m \geq p(0, \dots, 0)$$

for every choice of  $x$  and  $(u_1, \dots, u_m)$  such that  $f_i(x) \leq u_i$  for  $i = 1, \dots, m$ . Assuming  $p(0, \dots, 0)$  is finite, this condition is equivalent to the condition that  $u_i^* \geq 0$  for  $i = 1, \dots, m$  and

$$f_0(x) + u_1^* f_1(x) + \dots + u_m^* f_m(x) \geq p(0, \dots, 0), \quad \forall x \in \mathbb{R}^n.$$

In other words, the equilibrium prices are the same as the non-negative Lagrange multipliers  $u_1^*, \dots, u_m^*$  such that the unconstrained infimum of  $f_0 + u_1^* f_1 + \dots + u_m^* f_m$  coincides with the infimum of  $f_0$  subject to the constraints  $f_1(x) \leq 0, \dots, f_m(x) \leq 0$ .

These reflections on the nature of a classical nonlinear program lead us to propose a new concept of a generalized nonlinear program as, not just a single minimization problem, but a minimization problem with a built-in class of perturbations. In such a program, one is to study not only the infimum in the problem corresponding to zero perturbation, but also the sensitivity of the infimum with respect to perturbations to neighboring problems. The Lagrange multipliers are to be the "equilibrium prices" for the perturbations.

Suppose that for each vector  $u \in R^m$  we are given a pair  $(S_u, F_u)$ , where  $S_u$  is a subset of  $R^n$  (possibly empty) and  $F_u$  is a function on  $S_u$  with values in  $[-\infty, +\infty]$ . The correspondence

$$F: u \rightarrow (S_u, F_u)$$

will be called a bifunction from  $R^m$  to  $R^n$ . A bifunction is to be regarded as a generalization of "multivalued mapping": the image of  $u$  under  $F$  is not just a set, but a set with a distinguished function attached to it. One can interpret the function  $F_u$  as assigning a relative value or cost  $(F_u)(x)$  to each element  $x$  of the set  $S_u$ .

For any bifunction  $F$  from  $R^m$  to  $R^n$ , we define a generalized program (P): minimize the function  $F_0$  on the set  $S_0$ . The problem is to include the local analysis of the properties of the function  $p = \inf F$  at  $u = 0$ , where

$$(\inf F)(u) = \inf \{ (Fu)(x) \mid x \in Su \}.$$

(By convention, an infimum is  $+\infty$  if the set over which it is taken is empty.) A vector  $x \in R^n$  will be called an optimal solution to (P) if  $(\inf F)(0)$  is finite and attained at  $x$ . If  $(\inf F)(0)$  is finite, we define a Lagrange multiplier vector for (P) to be a vector  $u^* \in R^m$  such that

$$(\inf F)(u) + \langle u^*, u \rangle \geq (\inf F)(0)$$

for every perturbation  $u \in R^m$ . (Here  $\langle \cdot, \cdot \rangle$  denotes the ordinary inner product of two numerical vectors.)

Under simple convexity assumptions on the bifunction  $F$ , a comprehensive duality theory is possible for such generalized programs, as will be explained below. A dual program  $(P^*)$  may be constructed which is of the same type, except that it involves maximization rather than minimization. The dual of the program  $(P^*)$  is in turn (P). The extrema in (P) and  $(P^*)$  are generally equal. The optimal solutions to  $(P^*)$  are generally the Lagrange multiplier vectors for (P), while the optimal solutions to (P) are the Lagrange multiplier vectors for  $(P^*)$ . The pairs of optimal solutions to (P) and  $(P^*)$  are the saddle-points of a certain Kuhn-Tucker function.

An intriguing mathematical feature of the theory to be explained is that it constitutes a new "convex algebra" closely parallel to linear algebra. The convex bifunction  $F$  plays a role analogous to that of a linear transformation. Duality is obtained by the construction of an adjoint bifunction  $F^*$  in terms of Fenchel's conjugacy correspondence. Whereas a linear transformation and its adjoint are related by a bilinear function, a convex bifunction and its adjoint are related by a convex-concave function, and the formula  $\langle Fu, x^* \rangle = \langle u, F^*x^* \rangle$  appears as an "inf=sup" theorem for a dual pair of programs. Minimax theory is associated with the "inverse" operation for bifunctions.

The results in this paper are based on the general theory of convex functions and especially on the very important notion of conjugacy due to Fenchel [17]. The elementary facts about convex functions are reviewed in §2. Further details can be found in the works of Fenchel, Brøndsted, Moreau and Rockafellar listed among the references.

The complete proofs of the new duality theorems and of the theorems about bifunctions are all contained in a forthcoming book [44]. Some of the main ideas have already appeared in other papers of the author, however. A perturbational approach to duality theory is given in [43] and [38]. The correspondence between concave-convex functions on  $R^m \times R^n$  and convex functions on  $R^{m+n}$  (here the graph functions of convex bifunctions) is established in [38]. A similar "convex algebra" for multivalued mappings has been developed in [36] and described in [37].

Some applications of Fenchel's theory to general nonlinear programming have also been described by Ghouila-Houri [2], Dennis [7], Dieter [8,9], Falk and Thrall [15], Karlin [23], and Whinston [46].

An excellent discussion of general Lagrange multipliers as "equilibrium prices" has been given by Gale [19] in the case of concave maximization problems depending outside parameters.

## 2. Convex functions and their conjugates.

The object of the finite-dimensional theory of convex functions is the study of pairs  $(C, f)$ , where  $C$  is a non-empty convex set in  $R^n$  and  $f$  is a real-valued convex function on  $C$ , i.e. a function from  $C$  to  $R$  satisfying

$$(2.1) \quad f((1-\lambda)x + \lambda y) \leq (1-\lambda)f(x) + \lambda f(y), \quad 0 < \lambda < 1,$$

for any  $x \in C$  and  $y \in C$ . There are technical advantages, however, in representing each such pair by a function which is defined on all of  $R^n$  but which may have infinite values, namely the function obtained by defining  $f(x)$  to be  $+\infty$  for  $x \notin C$ .

In general, let  $f$  be any function defined on all of  $R^n$  and having values which are real numbers or  $+\infty$ . The epigraph of  $f$ , denoted by  $\text{epi } f$ , is the set of pairs  $(x, \mu)$  in  $R^{n+1}$  such that  $x \in R^n$ ,  $\mu \in R$  and  $\mu \geq f(x)$ . (Thus  $\text{epi } f$  can be regarded as the set of all "finite" points lying on or above the graph of  $f$ .) We define  $f$  to be a convex function on  $R^n$  if  $\text{epi } f$  is convex as a

subset of  $R^{n+1}$ . If there is no  $x$  such that  $f(x) = -\infty$ , this definition of convexity is equivalent to inequality (2.1) being satisfied throughout  $R^n$  (with the obvious rules for manipulating  $+\infty$ ).

If  $f$  is convex, the set

$$\text{dom } f = \{x \mid f(x) < +\infty\},$$

which is the projection of  $\text{epi } f$  on  $R^n$ , is convex; it is called the effective domain of  $f$ . A convex function  $f$  on  $R^n$  is said to be proper if  $\text{dom } f$  is non-empty and  $f$  is finite on  $\text{dom } f$ , in other words if  $f$  is not the constant function  $+\infty$  and there is no  $x$  such that  $f(x) = -\infty$ . The restriction of  $f$  to  $C = \text{dom } f$  is then a pair  $(C, f)$  of the type mentioned above, and every such pair arises in this way. Thus the study of the pairs  $(C, f)$  is replaced by the study of the proper convex functions  $f$  on  $R^n$ .

Convex functions which are not proper can arise naturally as the result of certain operations, and they do have some technical uses. The fundamental fact about an improper convex function  $f$  on  $R^n$  is that  $f$  must be identically  $-\infty$  on the interior of  $\text{dom } f$ .

A useful example of a convex function is the indicator function  $\delta(\cdot \mid C)$  of a convex set  $C$  in  $R^n$ , where  $\delta(x \mid C) = 0$  for  $x \in C$  and  $\delta(x \mid C) = +\infty$  if  $x \notin C$ . If  $f_0$  is a finite convex function on  $R^n$ , the convex function  $f = f_0 + \delta(\cdot \mid C)$

agrees with  $f_0$  on  $C$  and is  $+\infty$  elsewhere. Minimizing  $f_0$  on  $C$  is equivalent to minimizing  $f$  over all of  $R^n$ . We shall use this device to re-express all constrained extremum problems as formally unconstrained problems.

Let  $f$  be a convex function on  $R^n$ , and let  $D$  denote the collection of all pairs  $(x^*, \mu^*)$  such that  $x^* \in R^n$ ,  $\mu^* \in R$  and

$$f(x) \geq \langle x, x^* \rangle - \mu^*, \quad \forall x \in R^n.$$

The pointwise supremum of the corresponding collection of affine functions  $h(x) = \langle x, x^* \rangle - \mu^*$  is called the closure of  $f$  and is denoted by  $cl f$ . Thus by definition

$$(2.2) \quad (cl f)(x) = \sup \{ \langle x, x^* \rangle - \mu^* \mid (x^*, \mu^*) \in D \}.$$

When  $cl f = f$ , one says that  $f$  is closed. If  $f$  is proper, it can be shown that the epigraph of  $cl f$  is simply the closure in  $R^{n+1}$  of the epigraph of  $f$ . Then  $cl f$  is a closed proper convex function on  $R^n$ , and

$$(2.3) \quad (cl f)(x) = \liminf_{y \rightarrow x} f(y), \quad \forall x \in R^n.$$

In particular, a proper convex function is closed if and only if it is lower-semicontinuous, i.e. has the property that the convex level set  $\{x \mid f(x) \leq \mu\}$  is closed in  $R^n$  for each real  $\mu$ .

For a proper convex function  $f$ ,  $(cl f)(x)$  must actually coincide with  $f(x)$  for every  $x$  in the interior of  $\text{dom } f$  or outside the closure of  $\text{dom } f$ . Thus  $f \rightarrow cl f$  may be regarded as a regularizing operation which simply redefines  $f$  at certain boundary points of its effective domain, so as to make  $f$  lower-semicontinuous. For an

improper convex function  $f$ ,  $\text{cl } f$  is the constant function  $-\infty$  or the constant function  $+\infty$ , depending on whether or not  $\text{dom } f$  is non-empty.

Fenchel's important notion of conjugacy is obtained by further consideration of the set  $D$  introduced above. Clearly  $D$  consists of the pairs  $(x^*, \mu^*)$  in  $\mathbb{R}^{n+1}$  such that  $\mu^* \geq f^*(x^*)$ , where

$$(2.4) \quad f^*(x^*) = \sup_x \{ \langle x, x^* \rangle - f(x) \}.$$

Thus  $D$  is the epigraph of a certain function  $f^*$  on  $\mathbb{R}^n$ . This  $f^*$  is called the conjugate of  $f$ .

It can be seen that  $f^*$  is a closed convex function on  $\mathbb{R}^n$ , proper if and only if  $f$  itself is proper. The conjugate  $f^{**}$  of  $f^*$  is in turn given by

$$f^{**}(x) = \sup_{x^*} \{ \langle x, x^* \rangle - f^*(x^*) \}.$$

But this supremum is the same as the supremum in (2.2). Thus  $f^{**} = \text{cl } f$ . In particular, if  $f$  is closed it is the conjugate of its conjugate  $f^*$ .

Conjugacy therefore defines a one-to-one symmetric correspondence in the class of all closed convex functions on  $\mathbb{R}^n$ .

As an example, the conjugate of the indicator function  $\delta(\cdot | C)$  of a convex set  $C$  in  $\mathbb{R}^n$  is given by

$$\delta^*(x^* | C) = \sup_x \{ \langle x, x^* \rangle - \delta(x | C) \} = \sup_{x \in C} \langle x, x^* \rangle.$$

The function  $\delta^*(\cdot | C)$  is called the support function of  $C$ .

A convex function  $f$  on  $\mathbb{R}^n$  is necessarily continuous on the interior of its effective domain. It is differentiable

almost everywhere on any open set where it is finite.

Assume that  $x$  is any point where  $f$  is finite.

The (one-sided) directional derivative

$$(2.5) \quad f'(x; y) = \lim_{\lambda \downarrow 0} \frac{f(x + \lambda y) - f(x)}{\lambda}$$

exists and is a convex function of  $y$  (possibly with the values  $\pm\infty$ ). Of course, if  $f$  is actually differentiable at  $x$ , we have

$$(2.6) \quad f'(x; y) = \langle \nabla f(x), y \rangle,$$

where  $\nabla f(x)$  is the gradient of  $f$  at  $x$ ,

$$\nabla f(x) = \left( \frac{\partial f}{\partial x_1}(x), \dots, \frac{\partial f}{\partial x_n}(x) \right).$$

If  $f$  is not differentiable at  $x$ , the directional derivatives can still be described in terms of "subgradients". A

subgradient of  $f$  at  $x$  is a vector  $x^* \in \mathbb{R}^n$  such that

$$f(z) \geq f(x) + \langle z - x, x^* \rangle, \quad \forall z \in \mathbb{R}^n.$$

The set of subgradients  $x^*$  at  $x$  is a certain closed convex (possibly empty) set denoted by  $\partial f(x)$ . The case where  $\partial f(x)$  consists of just one  $x^*$  is precisely the case where  $f$  is finite and differentiable at  $x$ , the unique subgradient then being  $\nabla f(x)$ . It can be shown that, if  $x$  is actually an interior point of  $\text{dom } f$ ,  $\partial f(x)$  is non-empty and compact, and

$$(2.7) \quad f'(x; y) = \max \{ \langle x^*, y \rangle \mid x^* \in \partial f(x) \} = \delta^*(y \mid \partial f(x))$$

for each  $y \in \mathbb{R}^n$ . In general,  $\partial f(x)$  is empty and only if  $f'(x; y) = -\infty$  for some  $y$ .

When  $\partial f(x)$  is non-empty, one necessarily has  $(cl f)(x) = f(x)$ . On the other hand, when  $(cl f)(x) = f(x)$  one has  $x^* \in \partial f(x)$  if and only if  $x \in \partial f^*(x^*)$ . Thus the multivalued mapping  $\partial f^*: x^* \rightarrow \partial f^*(x^*)$  is the inverse of the multivalued mapping  $\partial f: x \rightarrow \partial f(x)$ , when  $f$  is a closed proper convex function.

Note that  $0 \in \partial f(x)$  if and only if  $f$  attains its minimum at  $x$ . We shall use this fact later in a slightly different form: when  $(cl f)(0) = f(0)$ , the vectors  $x^*$  in  $\partial f(0)$  are the same as those for which  $0 \in \partial f^*(x^*)$ , i.e. for which  $f^*$  attains its minimum.

The conjugate of a differentiable convex function  $f$  on  $R^n$  is closely related to the Legendre transform of  $f$ . Let  $C^*$  be the set of all gradients  $x^*$  of  $f$ , i.e. the image of  $R^n$  under the mapping  $x \rightarrow \nabla f(x)$ . Given any  $x^* \in C^*$ , the vectors  $x$  for which the supremum in (2.4) is attained are precisely those for which  $x^* = \nabla f(x)$ ; thus

$$(2.8) \quad f^*(x^*) = \langle x, x^* \rangle - f(x) \quad \text{when } x^* = \nabla f(x).$$

If the mapping  $\nabla f$  is one-to-one, we get

$$(2.9) \quad f^*(x^*) = \langle (\nabla f)^{-1}(x^*), x^* \rangle - f((\nabla f)^{-1}(x^*)), \quad x^* \in C^*.$$

This is the formula for the Legendre transform of  $f$ .

If  $\nabla f$  is not one-to-one, we can still conceive of parameterizing  $C^*$  in terms of  $x$  by means of the nonlinear substitution  $x^* = \nabla f(x)$ ; the substitution yields the formula

$$(2.10) \quad f^*(\nabla f(x)) = \langle x, \nabla f(x) \rangle - f(x).$$

This function of  $x$  is one which is common in the literature of nonlinear programming. It is generally not convex, of course, and it generally does not express  $f^*$  completely.

The set  $C^*$  need not be convex in  $R^n$ , and there may be points outside of  $C^*$  where  $f^*$  is finite and the Legendre transform is undefined.

It will be convenient in what follows to place concave functions on an equal footing with convex functions. A function  $g$  from  $R^n$  to  $[-\infty, +\infty]$  is said to be concave, of course, if  $f = -g$  is convex. All the above facts and definitions for convex functions have obvious analogues for concave functions, in which the roles of  $+\infty$ ,  $\inf$  and  $\leq$  are interchanged with those of  $-\infty$ ,  $\sup$  and  $\geq$ . In particular, the conjugate of a concave function  $g$  is defined by

$$g^*(x^*) = \inf_x \{ \langle x, x^* \rangle - g(x) \}.$$

It should be noted that  $g^*$  is not the same as  $-f^*$ , where  $f = -g$ . Instead one has  $g^*(x^*) = -f^*(-x^*)$ .

### 3. Dual programs and adjoint bifunctions.

By a convex bifunction from  $R^m$  onto  $R^n$ , we shall mean a correspondence  $F$  which assigns to each  $u \in R^m$  a function  $Fu$  from  $R^n$  to  $[-\infty, +\infty]$ , such that  $(Fu)(x)$  is a (jointly) convex function of  $(u, x)$  on  $R^{m+n}$ . This function on  $R^{m+n}$  is called the graph function of  $F$ . We shall say  $F$  is closed or proper according to whether its graph function is closed or proper. The effective domain of  $F$  is defined to be the (convex) projection on  $R^m$  of the effective domain of the graph function, i.e.

$$\text{dom } F = \{u \mid \exists x, (Fu)(x) < +\infty\}.$$

If  $F$  is closed, proper and convex, then in particular  $Fu$  is a closed convex function on  $R^n$  for every  $u$ , proper for  $u \in \text{dom } F$  but identically  $+\infty$  for  $u \notin \text{dom } F$ .

For example, let  $f_0, f_1, \dots, f_m$  be finite convex functions on  $R^n$ , and for each  $u = (u_1, \dots, u_m)$  define the function  $Fu$  by

$$(3.1) \quad (Fu)(x) = \begin{cases} f_0(x) & \text{if } f_1(x) \leq u_1, \dots, f_m(x) \leq u_m, \\ +\infty & \text{if not.} \end{cases}$$

It is easily demonstrated that  $F$  is a closed proper convex bifunction. Note that  $\text{dom } F$  consists of the vectors  $u$  such that the corresponding inequality system

$$f_1(x) \leq u_1, \dots, f_m(x) \leq u_m$$

has at least one solution  $x$ .

For another example, let  $A$  be a linear transformation from  $R^m$  to  $R^n$  and let

$$(3.2) \quad (Fu)(x) = \begin{cases} 0 & \text{if } x = Au, \\ +\infty & \text{if } x \neq Au. \end{cases}$$

This  $F$  is a closed proper convex bifunction which we call the indicator bifunction of  $A$ . We shall see that the "convex algebra" below reduces to ordinary linear algebra when the bifunctions are taken to be such indicator bifunctions.

Henceforth we assume for simplicity that  $F$  is a certain closed proper convex bifunction from  $R^m$  onto  $R^n$ .

The program  $(P)$  associated with  $F$ , as in the introduction, is that of minimizing  $FO$  on  $R^n$ . Of course, minimizing  $FO$  on  $R^n$  is equivalent to minimizing  $FO$  over the convex set  $\text{dom } (FO)$ , since  $FO$  has only the

value  $+\infty$  outside this set. The elements of  $\text{dom } (F)$  will be called the feasible solutions to (P). This is suggested by the case of (P) where  $F$  is given by (3.1), which we refer to as the case of a classical convex program. Feasible solutions to (F) exist if and only if  $0 \in \text{dom } F$ , in which event we say (P) is consistent. If  $0$  is actually an interior point of  $\text{dom } F$ , we say (P) is strictly consistent. In the classical case, (P) is strictly consistent if and only if there exists an  $x$  such that  $f_i(x) < 0$  for  $i = 1, \dots, m$ .

The fundamental and easily proved fact on which our analysis of (P) depends is that the function  $\inf F$ , where

$$(\inf F)(u) = \inf(Fu) = \inf_x (Fu)(x),$$

is a convex function on  $R^m$  whose effective domain is the same as  $\text{dom } F$ . The theory of closures, conjugates, directional derivatives and subgradients of convex functions can therefore be applied to the study of  $\inf F$  at  $u = 0$ .

For example, if (P) is strictly consistent,  $0$  is in the interior of the effective domain of  $\inf F$ , so we may conclude at once that  $(\inf F)(u)$  depends continuously on  $u$  for sufficiently small perturbations  $u$ .

Assume that  $(\inf F)(0)$  is finite. By definition,  $u^*$  is a Lagrange multiplier vector for (P) if and only if

$$(\inf F)(u) \geq (\inf F)(0) - \langle u, u^* \rangle, \quad \forall u \in R^m,$$

in other words if  $-u^*$  is a subgradient at  $0$ :

$$(3.3) \quad -u^* \in \partial(\inf F)(0).$$

If (P) is strictly consistent, so that 0 is an interior point of  $\text{dom}(\inf F)$ , we know from the general theory that  $\partial(\inf F)(0)$  is a non-empty compact convex set in  $R^m$  whose support function is the directional derivative function

$$(3.4) \quad (\inf F)'(0; u) = \lim_{\lambda \downarrow 0} \frac{(\inf F)(\lambda u) - (\inf F)(0)}{\lambda}.$$

In particular, a Lagrange multiplier vector  $u^* = (u_1^*, \dots, u_m^*)$  does exist when (P) is strictly consistent. This  $u^*$  is unique if and only if  $\inf F$  is actually differentiable at 0, in which case one has

$$(3.5) \quad u_i^* = -\frac{\partial}{\partial u_i} (\inf F)(0), \quad i = 1, \dots, m.$$

(Thus, for example, in a classical convex program the Lagrange multipliers, if unique, give the rates of change of the infimum with respect to changes of the constant terms in the corresponding constraint inequalities.) By the general theory of subgradients, a Lagrange multiplier vector fails to exist for (P) if and only if there exists a  $u$  such that  $(\inf F)'(0; u) = -\infty$ . The interpretation of this case is that there is some direction of perturbation in which the "minimal cost" drops off infinitely steeply, so that no finite "prices" for the perturbation variables can bring about a state of equilibrium.

To get the program dual to (P), we need to introduce the adjoint of  $F$ . This is the bifunction  $F^*$  from  $R^n$  onto  $R^m$  given by  $x^* \rightarrow F^*x^*$ , where

$$(3.6) \quad (F^*x^*)(u^*) = \inf_{u,x} \{ (Fu)(x) - \langle x, x^* \rangle + \langle u, u^* \rangle \}.$$

Note that, for the graph function  $f$  of  $F$ , one has

$$(F^*x^*)(u^*) = -\sup_{u,x} \{ \langle u, -u^* \rangle + \langle x, x^* \rangle - f(u, x) \} = -f^*(-u^*, x^*),$$

where  $f^*$  is the conjugate of  $f$  on  $R^{m+n}$ . Thus  $F^*$  is a closed proper concave bifunction in the obvious sense.

The adjoint of a concave bifunction is defined in the same way, except of course that "sup" replaces "inf".

Thus the adjoint  $F^{**}$  of  $F^*$  is defined in turn by

$$\begin{aligned} (F^{**}u)(x) &= \sup_{x^*, u^*} \{ (F^*x^*)(u^*) - \langle u, u^* \rangle + \langle x, x^* \rangle \} \\ &= \sup_{u^*, x^*} \{ \langle u, u^* \rangle + \langle x, x^* \rangle - f^*(u^*, x^*) \} = f^{**}(u, x). \end{aligned}$$

Since  $f^{**} = f$  under the conjugacy correspondence, we have  $F^{**} = F$ .

It is easy to see that, when  $F$  is the convex indicator bifunction of a linear transformation  $A$  from  $R^m$  to  $R^n$ ,  $F^*$  is the concave indicator bifunction of the adjoint linear transformation  $A^*$  from  $R^n$  back to  $R^m$  (corresponding to the transpose matrix), i.e.  $(F^*x^*)(u^*)$  is 0 if  $u^* = A^* x^*$  and  $-\infty$  if  $u^* \neq A^* x^*$ . In this sense, the adjoint operation for bifunctions generalizes the one for linear transformations. Further justification of the "adjoint" terminology will be given in the next section.

We define the dual program  $(P^*)$  to be that of maximizing the concave function  $F^*0$  on  $R^m$ . In  $(P^*)$  we are also interested in the properties of the function

$\sup F^*$  at  $x^* = 0$ , where

$$(\sup F^*)(x^*) = \sup (F^*x^*) = \sup_{u^*} (F^*x^*)(u^*).$$

Thus  $x^*$  is taken to be the perturbation variable in  $(P^*)$ , while  $u^*$  is the vector variable over which one maximizes.

Of course,  $\sup F^*$  turns out to be a concave function on  $R^n$ . All of what we have just said about  $\inf F$  in  $(P)$  applies to  $\sup F^*$  in  $(P^*)$  with only the obvious changes. The dual of  $(P^*)$  is in turn  $(P)$ , inasmuch as  $F^{**} = F$ .

As an example, let  $A$  be a linear transformation from  $R^n$  to  $R^m$ , fix  $a \in R^m$  and  $a^* \in R^n$ , and set

$$(3.7) \quad (Fu)(x) = \begin{cases} \langle x, a^* \rangle & \text{if } x \geq 0 \text{ and } Ax \geq a - u, \\ +\infty & \text{if not.} \end{cases}$$

(This is the case of (3.1) where the functions  $f_i$  are all affine.) Minimizing  $FO$  in  $(P)$  is then the same as minimizing  $\langle x, a^* \rangle$  subject to  $x \geq 0$  and  $Ax \geq a$ , an ordinary linear program. By a straightforward calculation from the definition of  $F^*$ ,

$$(3.8) \quad (F^*x^*)(u^*) = \begin{cases} \langle a, u^* \rangle & \text{if } u^* \geq 0 \text{ and } A^*u^* \leq a^* - x^*, \\ -\infty & \text{if not.} \end{cases}$$

Thus maximizing  $F^*0$  in  $(P^*)$  is the same as maximizing  $\langle a, u^* \rangle$  subject to  $u^* \geq 0$  and  $A^*u^* \leq a^*$ , the ordinary dual linear program.

The dual programs of Fenchel, extended by the present author in [43], may also be represented here as a special case. Again let  $A$  be a linear transformation from  $R^n$  to  $R^m$ , let  $f$  be a closed proper convex function on  $R^n$  and let  $g$  be a closed proper concave function on  $R^m$ .

Define  $F$  by

$$(3.9) \quad (Fu)(x) = f(x) - g(Ax + u).$$

Then  $F$  is a closed proper convex bifunction, and  $(P)$  consists of minimizing  $f(x) - g(Ax)$  in  $x \in R^n$ . Note that the perturbation  $u$  here corresponds to a translation of the function  $g$  on  $R^m$ . By elementary calculation,

$$(3.10) \quad (F^*x^*)(u^*) = g^*(u^*) - f^*(A^*u^* + x^*),$$

so that  $(P^*)$  consists of maximizing  $g^*(u^*) - f^*(A^*u^*)$  in  $u^* \in R^m$ . Fenchel's original programs are the case where  $A$  is the identity transformation.

For the classical convex program, the adjoint bifunction is given by

$$(F^*x^*)(u^*) = \begin{cases} -(f_0 + u_1^* f_1 + \dots + u_m^* f_m)^*(x^*) & \text{if } u^* = (u_1^*, \dots, u_m^*) \geq 0, \\ -\infty & \text{if } u^* \not\geq 0. \end{cases}$$

Thus the dual program  $(P^*)$  is to maximize  $-(f_0 + u_1^* f_1 + \dots + u_m^* f_m)^*(0)$  subject to  $u_i^* \geq 0, i = 1, \dots, m$ . To calculate the conjugate of  $f = f_0 + u_1^* f_1 + \dots + u_m^* f_m$  explicitly, one would have to know more about the given functions  $f_i$ . However, if every  $f_i$  is differentiable one can apply the Legendre transformation in the weakened form of (2.10) to  $f$  to get a problem which is "almost" equivalent to  $(P^*)$ . Since  $-f^*(\nabla f(x)) = f(x)$  by (2.10) when  $\nabla f(x) = 0$ , and

$$\nabla f = \nabla f_0 + u_1^* \nabla f_1 + \dots + u_m^* \nabla f_m,$$

the "approximate" problem is that of minimizing

$$f_0(x) + u_1^* f_1(x) + \dots + u_m^* f_m(x)$$

in  $u^* \in R^m$  and  $x \in R^n$  subject to the constraints

$$u^* \geq 0, \nabla f_0(x) + u_1^* \nabla f_1(x) + \dots + u_m^* \nabla f_m(x) = 0.$$

This is the well-known dual problem which was discovered by Wolfe [47].

It should be pointed out that the classical convex program can be modified in many ways by introducing additional perturbations. For instance, one can perturb the constraint  $f_i(x) \leq u_i$  by a translation  $y_i$  to the constraint  $f_i(x-y_i) \leq u_i$ . The dual problem would then turn out to involve an additional Lagrange multiplier vector  $y_i^*$  dual to the perturbation variable  $y_i \in R^n$ ; this would essentially be the dual problem for the classical convex program given by the author in [40]. The possibilities for perturbation are endless. The perturbations can be chosen to suit the situation, according to what "equilibrium prices" one is interested in. To apply the duality theory described here, it is only necessary that the perturbations be "convex", in the sense that the dependence of the problem on the perturbations be representable in terms of a convex bifunction  $F$ .

All the results relating the general dual pair of programs  $(P)$  and  $(P^*)$  are based on one elementary fact, which follows directly from the definitions: the convex minimand  $FO$  in  $(P)$  is the conjugate of the convex function  $-\sup F^*$  on  $R^n$ , while the concave maximand  $F^*O$  in  $(P^*)$  is the conjugate of the concave function  $-\inf F$  on  $R^m$ . This implies that

$$(FO)^* = (-\sup F^*)^{**} = -cl(\sup F^*),$$

$$(F^*O)^* = (-\inf F)^{**} = -cl(\inf F),$$

and hence that

$$(3.12) \quad \begin{aligned} \text{cl}(\sup F^*)(0) &= -\sup_x \{ \langle x, 0 \rangle - (F0)(x) \} = (\inf F)(0), \\ \text{cl}(\inf F)(0) &= -\inf_{u^*} \{ \langle 0, u^* \rangle - (F^*0)(u^*) \} = (\sup F^*)(0). \end{aligned}$$

The infimum  $(\inf F)(0)$  in  $(P)$  is thus always greater than or equal to the supremum  $(\sup F^*)(0)$  in  $(P^*)$ , and any possible discrepancy between these extrema is completely explained in terms of the closure operations for convex and concave functions.

Let us call  $(P)$  normal if  $\text{cl}(\inf F)(0) = (\inf F)(0)$ . If  $(P)$  is consistent, this is equivalent to the semicontinuity condition that

$$\lim_{u \rightarrow 0} \inf (\inf F)(u) = (\inf F)(0).$$

Similarly, let us call  $(P^*)$  normal if  $\text{cl}(\sup F^*)(0) = (\sup F^*)(0)$  in the sense of the closure operation for concave functions. Formulas (3.12) then yield a good duality theorem:  $(P)$  is normal if and only if  $(P^*)$  is normal. Moreover the normal case is precisely the one where the extrema in  $(P)$  and  $(P^*)$  are equal, i.e.

$$(3.13) \quad (\inf F)(0) = (\sup F^*)(0).$$

For brevity, we shall say that normality holds when both programs are normal and the "inf" and "sup" are equal. Normality holds in particular, then, when  $(P)$  is strictly consistent (since then  $\inf F$  is continuous at 0), or when a Lagrange multiplier vector exists for  $(P)$  (since then  $\partial(\inf F)(0) \neq \emptyset$ , implying that  $\text{cl}(\inf F)$  agrees with  $\inf F$  at 0). Likewise, normality holds when  $(P^*)$  is

strictly consistent, or when a Lagrange multiplier vector exists for  $(P^*)$ .

Suppose that normality holds, and that the common extremum value in (3.13) is finite. As we have already pointed out,  $u^*$  is a Lagrange multiplier vector for  $(P)$  if and only if  $u^* \in \partial(-\inf F)(0)$ . Since  $(-\inf F)^* = F^*0$ , this is equivalent to the condition that  $0 \in \partial(F^*0)(u^*)$ , i.e. that the concave function  $F^*0$  attain its maximum at  $u^*$ . Similarly, the Lagrange multiplier vectors  $x$  for  $(P^*)$  are the vectors where the convex function  $F0$  attains its minimum. This gives us another duality theorem:

assuming that normality holds, the Lagrange multiplier vectors  $u^*$  for  $(P)$  are precisely the optimal solutions (if any) to  $(P^*)$ , while the optimal solutions  $x$  to  $(P)$  are precisely the Lagrange multiplier vectors for  $(P^*)$ .

This type of duality has previously been known only in the linear programming case.

#### 4. Kuhn-Tucker functions and minimax theory.

We shall now describe the correspondence between convex bifunctions from  $R^m$  to  $R^n$  and concave-convex functions on  $R^m \times R^n$  which is analogous to the correspondence between linear transformations from  $R^m$  to  $R^n$  and bilinear functions  $R^m \times R^n$ . This correspondence gives further insight into the nature of the adjoint bifunction. It enables us to construct for each dual pair of programs  $(P)$  and  $(P^*)$  as in the last section a certain convex-concave function whose saddle-points correspond to optimal solutions to the programs much as in the

classical Kuhn-Tucker theory [24].

Let  $K$  be a concave-convex function on  $R^m \times R^n$ , i.e. a function with values in  $[-\infty, +\infty]$  such that  $K(u, v)$  is concave in  $u$  for each  $v$  and convex in  $v$  for each  $u$ . Closure operations may be applied to  $K$  for the sake of regularization. Let  $cl_v K$  be the function on  $R^m \times R^n$  obtained by closing  $K(u, v)$  as a convex function of  $v$  for each  $u$ . Similarly let  $cl_u K$  denote the function obtained by closing  $K$  as a concave function of  $u$  for each  $v$ . Then  $cl_u K$  and  $cl_v K$  are concave-convex functions on  $R^m \times R^n$  too.

We can proceed now to form the concave-convex functions  $cl_v cl_u K$  and  $cl_u cl_v K$ . The first of these is called the lower closure of  $K$  (since the final regularization involves lower-semicontinuity), and the second is called the upper closure of  $K$ . If  $K$  coincides with its lower closure, it is said to be lower-closed, and so forth. It turns out that  $cl_v cl_u K$  is itself always lower-closed, and  $cl_u cl_v K$  is upper-closed, but these two functions may disagree at certain points of  $R^m \times R^n$ .

Since the operations  $cl_v cl_u$  and  $cl_u cl_v$  do not quite produce the same result, there is not a unique natural closure operation for concave-convex functions. Nevertheless, there is an important phenomenon of pairing of closures. It may be shown that, if  $\underline{K}$  is any lower-closed concave-convex function on  $R^m \times R^n$ , then  $\bar{K} = cl_u \underline{K}$  is an upper-closed concave-convex function such that  $cl_v \bar{K} = \underline{K}$ .

Thus there is a simple one-to-one correspondence between the lower-closed functions and the upper-closed functions. Corresponding functions  $\underline{K}$  and  $\bar{K}$  generally have the same values, except at certain points, and  $\underline{K} \leq \bar{K}$ .

For example, let  $C$  and  $D$  be closed convex sets in  $R^m$  and  $R^n$ , respectively, and let  $K$  be any continuous finite concave-convex function defined on  $C \times D$ . Set

$$(4.1) \quad \begin{aligned} \underline{K}(u,v) &= \begin{cases} K(u,v) & \text{if } u \in C \text{ and } v \in D, \\ +\infty & \text{if } u \in C \text{ and } v \notin D, \\ -\infty & \text{if } u \notin C, \end{cases} \\ \bar{K}(u,v) &= \begin{cases} K(u,v) & \text{if } u \in C \text{ and } v \in D, \\ +\infty & \text{if } v \notin D \\ -\infty & \text{if } u \notin C \text{ and } v \in D. \end{cases} \end{aligned}$$

Then  $\underline{K}$  and  $\bar{K}$  are lower-closed and upper-closed concave-convex functions, respectively, which are paired together in the manner just described. Observe, incidentally, that

$$\begin{aligned} \sup_u \inf_v \underline{K}(u,v) &= \sup_u \inf_v \bar{K}(u,v) = \sup_{u \in C} \inf_{v \in D} K(u,v), \\ \inf_u \sup_v \underline{K}(u,v) &= \inf_u \sup_v \bar{K}(u,v) = \inf_{v \in D} \sup_{u \in C} K(u,v). \end{aligned}$$

Thus the minimax analysis of  $K$  with respect to  $C \times D$  can be represented by the formally unconstrained minimax analysis of  $\underline{K}$  or of  $\bar{K}$  (or of any extension of  $K$  to all of  $R^m \times R^n$  such that  $\underline{K} \leq K \leq \bar{K}$ ).

In order to apply these facts to the study of bifunctions in a manner suggestive of linear algebra, we introduce an inner product notation for the conjugate of a convex

(or concave) function  $f$ :

$$\langle f, x^* \rangle = \langle x^*, f \rangle = f^*(x^*).$$

Then, for any convex bifunction from  $R^m$  onto  $R^n$ , we can form

$$(4.2) \quad \langle Fu, x^* \rangle = \langle x^*, Fu \rangle = (Fu)^*(x^*)$$

as a function of  $u \in R^m$  and  $x^* \in R^n$ . Note that, if  $F$  is the indicator bifunction of a linear transformation  $A: R^m \rightarrow R^n$  as in (3.1), then  $\langle Fu, x^* \rangle$  is simply the bilinear function  $\langle Au, x^* \rangle$  associated with  $A$ .

The basic theorem is the following. If  $F$  is any closed convex bifunction from  $R^m$  onto  $R^n$ , then  $\langle Fu, x^* \rangle$  is a lower-closed concave-convex function on  $R^m \times R^n$ . Conversely, given any function  $\underline{K}$  of the latter type, there exists a unique closed convex bifunction  $F$  from  $R^m$  onto  $R^n$  such that  $\underline{K}(u, x^*) = \langle Fu, x^* \rangle$ , namely the  $F$  given by

$$(Fu)(x) = \sup_{x^*} \{ \langle x, x^* \rangle - \underline{K}(u, x^*) \}.$$

The upper-closed  $\bar{K}$  on  $R^m \times R^n$  paired with  $\underline{K}$  is precisely the concave-convex function associated with the adjoint bifunction  $F^*$ , i.e.

$$\bar{K}(u, x^*) = \langle u, F^*x^* \rangle = (F^*x^*)(u).$$

Thus the formulas

$$(4.3) \quad \begin{aligned} \text{cl}_u \langle Fu, x^* \rangle &= \langle u, F^*x^* \rangle, \\ \langle Fu, x^* \rangle &= \text{cl}_{x^*} \langle u, F^*x^* \rangle, \end{aligned}$$

hold for any closed convex bifunction and its adjoint.

Formulas (4.3) generalize the familiar formula

$$\langle Au, x^* \rangle = \langle u, A^*x^* \rangle$$

relating a linear transformation and its adjoint. Since the closure operations in (4.3) merely redefine the functions

at special points, one will actually have

$$(4.4) \quad \langle Fu, x^* \rangle = \langle u, F^*x^* \rangle$$

for "most" values of  $u$  and  $x^*$ .

Observe that (4.4) expresses a duality between two different extremum problems, because by definition

$$(4.5) \quad \begin{aligned} \langle Fu, x^* \rangle &= \sup_x \{ \langle x, x^* \rangle - (Fu)(x) \} \\ \langle u, F^*x^* \rangle &= \inf_{u^*} \{ \langle u, u^* \rangle - (F^*x^*)(u^*) \} . \end{aligned}$$

In particular, we have

$$(4.6) \quad \begin{aligned} -\langle Fu, 0 \rangle &= \inf_x (Fu)(x) = (\inf F)(u), \\ -\langle 0, F^*x^* \rangle &= \sup_{u^*} (F^*x^*)(u^*) = (\sup F^*)(x^*). \end{aligned}$$

The equality of the extrema in the programs (P) and (P\*) in the last section is therefore expressed simply by  $\langle FC, 0 \rangle = \langle 0, F^*0 \rangle$ .

Minimax characterizations of duality are obtained through the introduction of inverse bifunctions. The inverse of a convex bifunction  $F$  from  $R^m$  onto  $R^n$  is the concave bifunction  $F_*$  from  $R^n$  to  $R^m$  defined by

$$(4.7) \quad (F_*x)(u) = -(Fu)(x).$$

If  $F$  is closed,  $F_*$  is closed too. The inverse of a concave bifunction is defined in the same way. It is easily seen that  $F_{**} = F$  and  $(F^*)_* = (F_*)^*$ . The latter bifunction from  $R^m$  to  $R^n$  will be denoted simply by  $F^\circ$ .

As an example, if  $F$  is the convex indicator bifunction of a non-singular linear transformation  $A$  as in (3.2), then  $F_*$  is the concave indicator bifunction of  $A^{-1}$ , i.e.  $(F_*x)(u)$  is 0 if  $u = A^{-1}x$  and  $-\infty$  if  $u \neq A^{-1}x$ . Likewise,  $F^\circ$

is the convex indicator bifunction of  $A^{*-1}$ .

Given any closed proper convex bifunction  $F$  from  $R^m$  to  $R^n$ , we define the Kuhn-Tucker function of the corresponding program (P) to be  $\langle u^*, F, x \rangle$  as a function of  $u^*$  and  $x$ . Since  $F, x$  is concave, we have by definition

$$(4.8) \quad \begin{aligned} \langle u^*, F, x \rangle &= \inf_u \{ \langle u, u^* \rangle - (F, x)(u) \} \\ &= \inf_u \{ \langle u, u^* \rangle + (Fu)(x) \}. \end{aligned}$$

This is, of course, an upper-closed concave-convex function on  $R^m \times R^n$  by the correspondence theory already outlined.

In the case of a classical convex program, where  $F$  is given by (3.1), the Kuhn-Tucker function is evidently given by

$$(4.9) \quad \langle u^*, F, x \rangle = \begin{cases} f_0(x) + u_1^* f_1(x) + \dots + u_m^* f_m(x) & \text{if } u^* = (u_1^*, \dots, u_m^*) \geq 0 \\ -\infty & \text{if } u^* \not\geq 0. \end{cases}$$

Except for the convenient concave extension by means of  $-\infty$ , this is the function associated with (P) by the familiar Kuhn-Tucker theory.

In the case where  $F$  is given by (3.9), the Kuhn-Tucker function is given by

$$(4.10) \quad \langle u^*, F, x \rangle = f(x) + g^*(u^*) - \langle Ax, u^* \rangle,$$

with  $-\infty + \infty$  taken to be  $+\infty$ .

A saddle-point of the Kuhn-Tucker function is, of course, a vector pair  $(u^*, x)$  such that

$$(4.11) \quad \langle u^{*'}, F, x \rangle \leq \langle u^*, F, x \rangle \leq \langle u^*, F, x' \rangle, \quad \forall u^{*'}, \forall x'.$$

The main result is this: a vector pair  $(u^*, x)$  is a saddle-point of the Kuhn-Tucker function of (P) if and only if  $u^*$  is a Lagrange multiplier vector for (P) and

$x$  is an optimal solution to (P). In this event normality holds, and the minimax value  $\langle u^*, F, x \rangle$  coincides with the infimum in (P) and the supremum in (P\*). Moreover, as explained in the last section,  $u^*$  is then dually an optimal solution to (P\*), and  $x$  is a Lagrange multiplier vector for (P\*).

Given any upper-closed concave-convex function  $\bar{K}$  on  $R^m \times R^n$  (for instance a  $\bar{K}$  of the type in (4.1)), there is, as we know, a unique closed concave bifunction  $G$  from  $R^n$  onto  $R^m$  such that  $\bar{K}(u^*, x) = \langle u^*, Gx \rangle$ . Hence there is a unique program (P) having  $\bar{K}$  as its Kuhn-Tucker function, namely the (P) corresponding to  $F = G$ . The inverse operation for bifunctions therefore corresponds to a general minimax theory for concave-convex functions in the same way that the adjoint operation for bifunctions corresponds to a general duality theory for convex programs. It is clear from the definitions that the  $F$  and  $F^*$  here are expressible in terms of  $\bar{K}$  by

$$(4.12) \quad \begin{aligned} (Fu)(x) &= \sup_{u^*} \{ \bar{K}(u^*, x) - \langle u, u^* \rangle \}, \\ (F^*x^*)(u^*) &= \inf_x \{ \bar{K}(u^*, x) - \langle x, x^* \rangle \}. \end{aligned}$$

In particular, the minimand in (P) is given by

$$(FO)(x) = \sup_{u^*} \bar{K}(u^*, x),$$

and the maximand in (P\*) is given by

$$(F^*O)(u^*) = \inf_x \bar{K}(u^*, x).$$

The dual programs of Dantzig, Eisenberg and Cottle [6], Stoer [45], Mangasarian and Ponstein [26], Falk and Thrall<sup>[5]</sup> may be obtained in this way, for instance by applying the Legendre

transformation formula (2.10) to (4.12) and similar devices.

The pair of functions  $\langle F; u; x \rangle$ ,  $\langle u; F; x \rangle$ , is conjugate to the pair of functions  $\langle Fu, x^* \rangle$ ,  $\langle u, F^*x^* \rangle$ , in the following sense. If  $K$  is any one of the concave-convex functions such that

$$(4.13) \quad \langle Fu, x^* \rangle \leq K(u, x^*) \leq \langle u, F^*x^* \rangle$$

(such functions all being essentially the same, except at special points), one has, according to the definitions,

$$(4.14) \quad \inf_u \sup_{x^*} \{ \langle u, u^* \rangle + \langle x, x^* \rangle - K(u, x^*) \} = \langle u^*, F; x \rangle, \\ \sup_{x^*} \inf_u \{ \langle u, u^* \rangle + \langle x, x^* \rangle - K(u, x^*) \} = \langle F; u^*, x \rangle.$$

On the other hand, if  $K^*$  is any one of the functions satisfying

$$(4.15) \quad \langle F; u^*, x \rangle \leq K^*(u^*, x) \leq \langle u^*, F; x \rangle,$$

one has in turn

$$(4.16) \quad \inf_{u^*} \sup_x \{ \langle u, u^* \rangle + \langle x, x^* \rangle - K^*(u^*, x) \} = \langle u, F^*x^* \rangle, \\ \sup_x \inf_{u^*} \{ \langle u, u^* \rangle + \langle x, x^* \rangle - K^*(u^*, x) \} = \langle Fu, x^* \rangle.$$

Applying (4.3) to the convex bifunction  $F$ : in place of  $F$ , we have

$$(4.17) \quad \text{cl}_{u^*} \langle F; u^*, x \rangle = \langle u^*, F; x \rangle, \\ \langle F; u^*, x \rangle = \text{cl}_x \langle u^*, F; x \rangle,$$

and this makes possible a detailed comparison of the "inf sup" in (4.14). In particular we see that these two extrema are "usually" equal; the fact that they can be different in some cases is exactly dual to the fact that the upper and lower closure operations for concave-convex functions do not always coincide. A minimax theory from this point of view was developed by the author in [35].

References

1. K.J. Arrow, L. Hurwicz and H. Uzawa,  
Studies in Linear and Nonlinear Programming (Stanford University Press, 1958).
2. C. Berge and A. Ghouila-Houri,  
Programmes, Jeux et Réseaux de Transport (Dunod, Paris, 1962).
3. A. Brøndsted,  
"Conjugate convex functions in topological vector spaces", Mat.-fys Medd. Dansk. Vid. Selsk. 34 (1964), No. 2, 1-26.
4. A. Charnes, W.W. Cooper and K. Kortanek,  
"A duality theory for convex programs with convex constraints", Proc. Amer. Math. Soc. 68 (1962), 605-608.
5. R.W. Cottle,  
"Symmetric dual quadratic programs", Q. Appl. Math. 21 (1963), 237.
6. G. Dantzig, E. Eisenberg and R.W. Cottle,  
"Symmetric dual nonlinear programs", Pacific J. Math. 15 (1965), 809-812.
7. J.B. Dennis,  
Mathematical Programming and Electrical Networks (Technology Press, Cambridge, Mass., 1959).
8. U. Dieter,  
"Dualität bei konvexen Optimierungs-(Programmierungs-) Aufgaben", Unternehmensforschung 9 (1965), 91-111.
9. U. Dieter,  
"Optimierungsaufgaben in topologischen Vektorräumen I: Dualitätstheorie", Z. Wahrscheinlichkeitstheorie verw. Geb. 2 (1966), 89-117.
10. R.J. Duffin,  
"Infinite programs", Annals of Math. Study 38 (1956), 157-170.
11. W.S. Dorn,  
"Duality in quadratic programming", Q. Appl. Math. 18 (1960), 155-162.
12. W.S. Dorn,  
"A duality theorem for convex programs", IBM J. Research Devel. 4 (1960), 407-413.

13. R.J. Duffin,  
"Dual programs and minimum cost", J. Soc. Indust.  
Appl. Math. 10 (1962), 119-124.
14. E. Eisenberg,  
"Duality in homogeneous programming", Proc. Amer.  
Math. Soc. 12 (1961), 783-787.
15. J.E. Falk and R.M. Thrall,  
"A constrained Lagrangian approach to nonlinear  
programming", technical report, (Dept. of Math.,  
Univ. of Michigan, Ann Arbor, 1965).
16. K. Fan, I. Glicksberg and A.J. Hoffman,  
"Systems of inequalities involving convex functions",  
Proc. Amer. Math. Soc. 27 (617-622).
17. W. Fenchel,  
"On conjugate convex functions", Canad. J. Math.  
1 (1949), 73-77.
18. W. Fenchel,  
Convex Cones, Sets and Functions, lecture notes,  
Princeton University, 1951.
19. D. Gale,  
"A geometric duality theorem with economic application",  
Review of Econ. Studies 34 (1967).
20. M.A. Hanson,  
"A duality theorem in nonlinear programming with  
nonlinear constraints", Austral. J. Statist. 2 (1961),  
64-72.
21. P. Huard,  
"Dual programs", IBM J. Research Devel. 6 (1962),  
137-139.
22. F. John,  
"Extremum problems with inequalities as subsidiary  
conditions", Studies and Essays, Courant Anniversary  
Volume (Interscience, New York, 1948).
23. S. Karlin,  
Mathematical Methods and Theory in Games, Programming  
and Economics (Mc Graw-Hill, New York, 1960).
24. H.W. Kuhn and A.W. Tucker,  
"Nonlinear programming", Proceedings of the Second  
Berkeley Symposium on Mathematical Statistics and  
Probability (Univ. of California Press, Berkeley,  
1951), 481-492.

25. O.L. Mangasarian,  
"Duality in nonlinear programming", Quart. Appl. Math. 20 (1962), 300-302.
26. O.L. Mangasarian and J. Ponstein,  
"Minimax and duality in nonlinear programming", J. Math. Anal. Appl. 11 (1965), 504-518.
27. J.-J. Moreau,  
"Fonctions convexes en dualité", multigraph, Seminaires de Math., Faculté des Sciences, Université de Montpellier (1962).
28. J.-J. Moreau,  
"Etude locale d'une fonctionnelle convexe", multigraph, Seminaires de Math., Faculté des Sciences, Université de Montpellier (1963).
29. J.-J. Moreau,  
"Fonctionnelles sous-differentiables", C.R. Acad. Sci. 257 (1963), 4117-4119.
30. J.-J. Moreau,  
"Théorèmes 'inf-sup'", C.R. Acad. Sci. 258 (1964), 2720-2722.
31. J.-J. Moreau,  
"Proximité et dualité dans un espace hilbertien", Bull. Soc. Math. France 93 (1965), 273-299.
32. J.-J. Moreau,  
Fonctionnelles Convexes, lecture notes, Seminaire "Equations aux dérivées partielles", Collège de France, 1966.
33. R.T. Rockafellar,  
Convex Functions and Dual Extremum Problems, thesis, Harvard, 1963.
34. R.T. Rockafellar,  
"Duality theorems for convex functions", Bull. Amer. Math. Soc. 70 (1964), 189-192.
35. R.T. Rockafellar,  
"Minimax theorems and conjugate saddle-functions", Math. Scand. 14 (1964), 151-173.
36. R.T. Rockafellar,  
Monotone Processes of Concave and Convex Type, (1965), being published as a Memoir of the Amer. Math. Soc.

37. R.T. Rockafellar,  
"A monotone convex analog of linear algebra",  
Proceedings of the Colloquium on Convexity, Copenhagen, 1965, W. Fenchel, ed. (Matematisk Institut, Copenhagen, 1967), 261-276.
38. R.T. Rockafellar,  
"A general correspondence between dual minimax problems and convex programs", (1965), to appear in Pacific J. Math.
39. R.T. Rockafellar,  
"Helly's theorem and minima of convex functions",  
Duke Math. J. 32 (1965), 381-398.
40. R.T. Rockafellar,  
"An extension of Fenchel's duality theorem for convex functions", Duke Math. J. 33 (1966), 81-90.
41. R.T. Rockafellar,  
"Convex analysis", lecture notes, Princeton University, 1966.
42. R.T. Rockafellar,  
"Conjugates and Legendre transforms of convex functions", Canad. J. Math. 19 (1967), 200-205.
43. R.T. Rockafellar,  
"Duality and stability in extremum problems involving convex functions", Pacific J. Math. 21 (1967), 167-187.
44. R.T. Rockafellar,  
Convex Analysis, in preparation for the Princeton University Press.
45. J. Stoer,  
"Über einen Dualitätssatz der nichtlinearen programmierung", Num. Math. 6 (1964), 55-58.
46. A. Whinston,  
"Some applications of the conjugate function theory to duality", Nonlinear Programming, J. Abadie, ed. (North-Holland, Amsterdam, 1967), 75-96.
47. P. Wolfe,  
"A duality theorem for nonlinear programming",  
Q. Appl. Math. 19 (1961), 239-244.

Department of Mathematics  
University of Washington  
Seattle, Washington 98105

**MATHEMATICAL ECONOMICS**

by

**KENNETH ARROW**

at the

**American Mathematical Society Summer Seminar**

on the

**Mathematics of the Decision Sciences**

**Stanford University**

**July - August 1967**

# APPLICATIONS OF CONTROL THEORY TO ECONOMIC GROWTH

Kenneth J. Arrow

## Lecture 1

### Review of the Basic Theorems of Optimal Control Theory: Finite Horizon

The basic criteria for optimization of dynamic processes in continuous times, as stated by L. S. Pontryagin and associates [1962] will be restated in this and the following lecture. Some emphasis will be placed on special features appropriate to the use that will be made of these theorems in growth theory, in particular the assumption of an infinite horizon and the presence of constraints on the choice of control variables.

The object of study is a system, economic or other, evolving in time. At any moment, the system is in some state, which can be described by a finite-dimensional vector  $x(t)$ . For an economic system, the amount of capital goods of each type might constitute a suitable state description.

In an optimization problem there is some possibility of controlling the system. At any time,  $t$ , there is a vector  $v(t)$  which can be chosen by a decision-maker from some set which may, in general, vary with both  $t$  and the state  $x(t)$ . The vector  $v(t)$  is frequently referred to as the decision or control variable; following the terminology of Tinbergen [1952, p. 7], the term instrument is used here. In an economic system the instruments are typically the allocations of resources to different productive uses and to consumption or perhaps taxes and bond issues which at least partially determine allocations.

It is assumed that the state and the instrument variables at any point of time completely determine the rate of change of the state of the system. Thus, for a given technology and labor force, the capital structure (state)

together with its allocation among different uses (by some of the instruments) determine the outputs of all goods. These in turn are allocated between consumption and capital accumulation (through other components of the instrument vector). In symbols, the evolution of the state of the system is governed by the differential equations

$$(1) \quad \dot{x} = T[x(t), v(t), t],$$

which will be referred to as transition equations. The time variable  $t$  may enter into  $T$  to allow for the possibility that the transition relations may vary over time due to technological progress, labor force growth or other exogenous factors.

Given, then, the state of the system at some time, say 0, and the choice of instruments as a function of time,  $v(t)$ , the whole course of the system is determined. To begin with, let us suppose that the analysis is carried out only until a finite horizon  $T$ , after which the process ceases.

By suitable choices of the values of instruments over time, alternative histories of the process can be achieved. As usual in economic analysis, we assume that these histories can be valued in some way, i.e., we can express preferences as between alternative histories, and these preferences can be given numerical value, a utility functional with arguments  $x(t)$ ,  $v(t)$ , ( $0 \leq t \leq T$ ). The optimization problem is to choose the values of the instrument variables so as to maximize the utility functional subject to the constraints implied by (1), the constraints on the choices of the instruments, and the initial values of the state variables.

More specifically, it will be assumed that the utility functional is additive over time. That is, at each moment  $t$  there is a return or felicity (to use a term due to Gorman [1957, p. 43]) which depends only on

the values of the state variables and instruments at time  $t$ , such that the utility of a whole history is the sum or integral of the values of the felicities at each moment of time. Let

- (2)  $U(x,v,t)$  = felicity at time  $t$  if the state is  $x$  and the instrument vector is  $v$ .

In addition to the felicity generated at each moment of time during the process, the decision-maker may also assign a value to the state achieved at the end of the process,  $T$ . In an industrial application the stock of machines may have a scrap value, and we will use this term generally. In a broader economic context, if  $T$  is not literally the end of the world but only the end of the planning period, the capital stock left over at  $T$  will have some use in the future. The scrap value will be denoted by  $S[x(T)]$ .

The general form of the optimization problem in time is then, with finite horizon,

- (3) Maximize  $\int_0^T U[x(t), v(t), t] dt + S[x(T)]$  with respect to choice of the instruments over time subject to (1), some constraints on the choice of instruments possibly depending on the current values of the state variables, and the initial conditions  $x(0) = x$ .

Then Pontryagin and associates [1962] (see also Halkin [1964]) have shown

Proposition 1. Let  $v^*(t)$  be a choice of instruments ( $0 \leq t \leq T$ ) which maximizes (3). Then there exist auxiliary variables, functions of time,  $p(t)$ , with the same dimensionality as the state  $x$ , such that, for each  $t$ ,

- (a)  $v^*(t)$  maximizes  $H[x(t), v(t), p(t), t]$ , where

$$H(x,v,p,t) = U(x,v,t) + pT(x,v,t);$$

the function  $p(t)$  satisfies the differential equations

$$(b) \quad \dot{p}_1 = -\partial H / \partial x_1, \text{ evaluated at } x = x(t), v = v^*(t), \\ p = p(t);$$

and the transversality conditions

$$(c) \quad p_1(T) = \partial S / \partial x_1, \text{ evaluated at } x = x(T),$$

hold.

The function  $H$  is known as the Hamiltonian. The auxiliary variables  $p$  can be given an economic interpretation: Consider the maximization of the utility function from any time  $t_0$  to the horizon  $T$ ; the past history, before  $t_0$ , affects this problem only through the state at time  $t_0$ , as can easily be seen from (1) and the additive nature of the maximand in (3). Let this maximum be

$$(4) \quad V(x, t_0) = \max \left\{ \int_{t_0}^T U[x(t), v(t), t] dt + S[x(T)] \right\}, \text{ where } x(t_0) = x$$

Then the auxiliary variables are defined so that

$$(5) \quad p_1 = \partial V / \partial x_1.$$

An auxiliary variable measures the marginal contribution of the corresponding state variable to the utility functional at time  $t_0$ . Then  $p_1 x_1 = p_1 T_1$  is the rate of increase of utility due to the current rate of increase of the state variable  $x_1$ , and therefore  $H$  is the current flow of utility from all sources, both enjoyed immediately, as expressed by  $U$ , and anticipated to be enjoyed in the future, as expressed by  $pT$ . The current instruments are chosen then to maximize  $H$ . The condition (b) is an equilibrium condition for holding the state variables constant (at an instant of time); the increment in utility plus speculative gain should be zero; if not, the individual would have wanted to have less or more of

that state variable (read, stock of a capital good in the economic context). Finally, at time  $T$ ,  $V(x, T) = S(x)$ ; hence (c) holds by (5).

In the sequel, a slightly different formulation of end-of-period conditions will be useful. Instead of a scrap value, require simply that the end-of-period values of the state variables be non-negative. Now approximate this condition by a scrap value function; that is, permit negative values but impose a very large penalty. Formally, let

$$S(x) = \sum_{i=1}^n P_i \min(x_i, 0),$$

where  $\min(x_i, 0)$  means the smaller of  $x_i$  and 0, and the  $P_i$ 's are chosen very large. For  $x_i < 0$ ,  $\partial S / \partial x_i = P_i$ ; for  $x_i > 0$ ,  $\partial S / \partial x_i = 0$ . If  $x_i = 0$ , the right-hand derivative is zero and the left-hand derivative  $P_i$ , a fact which may be expressed loosely by the statement  $0 \leq \partial S / \partial x_i \leq P_i$ .

Now let the  $P_i$ 's approach  $+\infty$ , so that we may be sure that the optimal policy will never lead to a final negative value,  $x_i(T)$ :

$$x(T) \geq 0.$$

From the preceding discussion,  $\partial S / \partial x_i \geq 0$ , and further if  $x_i(T) > 0$ , then  $\partial S / \partial x_i = 0$ . In view of Proposition 1(c),  $p_i(T) \geq 0$ ,  $p_i(T) x_i(T) = 0$ , all  $i$ , or

$$p(T) \geq 0, \quad p(T) x(T) = 0.$$

Proposition 2. Let  $v^*(t)$  be a choice of instruments ( $0 \leq t \leq T$ ) which maximizes  $\int_0^T U[x(t), v(t), t] dt$  subject to the conditions

$$(a) \quad \dot{x} = T[x(t), v(t), t],$$

some constraints on the choices of instruments possibly involving current values of the state variables, and the terminal conditions,  $x(T) \geq 0$ . Then there exist auxiliary variables,  $v(t)$ , such that (a) and (b) of Proposition 1

hold and for which

$$(b) \quad p(T) \geq 0, \quad p(T) x(T) = 0.$$

The optimal path is the solution of the differential equations (1) and Proposition 1(b); the values of the instruments which enter into them are determined as functions of  $x$ ,  $p$ , and  $t$  by Proposition 1(a). The number of these equations is twice that of the number of state variables. The solution is usually only determined when an equal number of initial conditions are specified. The values of the state variables at the beginning of the process,  $x(0)$ , are taken as known, but these constitute only half the needed conditions. The transversality conditions, Proposition 1(c) or Proposition 2(b), constitute the remaining conditions, but from a practical point of view they suffer from the severe difficulty of being defined at the end of the process, while the other initial conditions are defined at the beginning. The computation can proceed by guessing initial values,  $p(0)$ , solving the system of transition and auxiliary equations with the hope that the transversality conditions are satisfied, and correcting the initial guesses if not. It can also proceed by guessing the final state  $x(T)$  and solving the equations backward in the hope that the initial conditions are satisfied.

Now consider more explicitly the constraints on the instruments. In general, they may depend on the values of the state variables. Thus, amounts of resources allocated to particular productive purposes are constrained by the total amounts available, which in turn are determined by the state variables. The following discussion is based on that in Pontryagin [1962, Chapter VI] and on the theory of nonlinear programming due to Kuhn and Tucker [1951].

Let the choice of instruments at any time  $t$  with state  $x$  satisfy a vector of inequality constraints,

$$(6) \quad F(x, v, t) \geq 0.$$

For example, if output is a function of the stock of capital,  $F(K)$ , and if output is to be divided between consumption ( $C$ ) and investment ( $I$ ), then the instruments  $C$  and  $I$  satisfy the condition,

$$F(K) - C - I \geq 0,$$

which involves the state variable  $K$ . Some of the constraints in  $F$  might not include state variables; for example, non-negativity conditions on the instruments.

It is well known from the general theory of nonlinear programming that if  $v^*$  maximizes  $H$  subject to the conditions (6), and if these constraints satisfy a certain condition known as the Constraint Qualification (see Kuhn and Tucker, [1951, pp. 483-4]; Arrow, Hurwicz, and Uzawa [1961]), then at any moment of time there exist multipliers  $q^0$  such that

$$(7) \quad q^0 \geq 0, \quad q^0 F'(x, v^*, t) = 0,$$

and

$$(8) \quad \partial L / \partial x_i = 0 \quad \text{at } v = v^*, \quad q = q^0,$$

where

$$(9) \quad L = H + qF$$

It can be shown that  $\partial H / \partial x_i = \partial L / \partial x_i$  when evaluated at  $v = v^*$ ,  $q = q^0$ .

With the explicit formulation (6) of constraints, the conditions for optimization over time become

Proposition 3. Let  $v^*(t)$  be a choice of instruments ( $0 \leq t \leq T$ ) which maximizes  $\int_0^T U[x(t), v(t), t] dt$  subject to the conditions,

$$(a) \quad \dot{x} = T[x(t), v(t), t],$$

a set of constraints,

$$(b) \quad F(x(t), v(t), t) \geq 0,$$

on the instruments possibly involving the state variables, initial conditions on the state variables, and the terminal conditions  $x(T) \geq 0$ . If the Constraint Qualification holds, then there exist auxiliary variables  $p(t)$ , such that, for each  $t$ ,

(c)  $v^*(t)$  maximizes  $H[x(t), v, p(t), t]$  subject to the constraints (b), where  $H(x, v, p, t) = U(x, v, t) + pT(x, v, t)$ ,

$$(d) \quad p_1 = -\partial L / \partial x_1, \text{ evaluated at } x = x(t), v = v^*(t), p = p(t),$$

where

$$(e) \quad L(x, v, q, t) = H(x, v, p, t) + qF(x, v, t),$$

and the Lagrange multipliers  $q$  are such that

$$(f) \quad \partial L / \partial v_k = 0, \text{ for } x = x(t), v = v^*(t), p = p(t),$$

$$q \geq 0, qF[x(t), v^*(t), t] = 0,$$

and

$$(g) \quad p(T) \geq 0, p(T) x(T) = 0.$$

In many circumstances it is reasonable to consider in addition restrictions on the state variables in which the instruments do not enter. In particular, if the state variables are stocks of capital, negative values have no meaning. Here, non-negativity conditions on the state variables,

$$(10) \quad x(t) \geq 0,$$

will be considered; the terminal condition  $x(T) \geq 0$  is implied.

For any  $i$ , if  $x_i(t) > 0$ , then the corresponding constraint (10) is ineffective and can be disregarded. Suppose that  $x_i(t) = 0$  over some interval. Then, to avoid violation of (10), the instruments must be so constrained that  $\dot{x}_i(t) \geq 0$ , and this constraint is clearly effective over that interval. But  $\dot{x}_i = T_i$ , so that the constraint  $T_i(x, v, t) \geq 0$  is effective over that interval. Then, in Proposition 3, this constraint can be regarded as added to the original set of constraints (b). Let  $q$  be the Lagrange multipliers associated with the original constraints (b), and let  $r_i$  be the multiplier associated with the new constraint  $T_i \geq 0$ . As before,  $r_i \geq 0$ . Define, in addition,  $r_i = 0$  for each state variable for which  $x_i(t) > 0$ . Then clearly  $r \geq 0$ ,  $rT = 0$ ,  $rx = 0$ .

Proposition 4. Let  $v^*(t)$  be a choice of instruments ( $0 \leq t \leq T$ ) which maximizes  $\int_0^T U[x(t), v(t), t] dt$  subject to the conditions,

$$(a) \quad \dot{x} = T[x(t), v(t), t],$$

a set of constraints,

$$(b) \quad F[x(t), v(t), t] \geq 0,$$

involving the instruments and possibly the state variables, initial conditions on the state variables, and the non-negativity conditions,

$$(c) \quad x(t) \geq 0,$$

on the state variables. If the Constraint Qualification holds, then there exist auxiliary variables  $p(t)$  such that, for each  $t$ ,

(d)  $v^*(t)$  maximizes  $H[x(t), v, p(t), t]$  subject to the constraints (b) and the additional constraints  $T_i[x(t), v, t] \geq 0$  for all  $i$  for which  $x_i(t) = 0$ , where  $H(x, v, p, t) = U(x, v, t) + pT(x, v, t)$ ;

(e)  $p = -\partial L / \partial x_i$ , evaluated at  $x = x(t)$ ,  $v = v^*(t)$ ,  $p = p(t)$ ,  $q = q(t)$ ,  $r = r(t)$ , where

$$(f) \quad L(x, v, p, q, r, t) = H(x, v, p, t) + qF(x, v, t) + rT(x, v, t),$$

and the Lagrange multipliers  $q$  and  $r$  are such that

$$(g) \quad \partial L / \partial v_k = 0, \text{ for } x = x(t), v = v^*(t), p = p(t),$$

$$q(t) \geq 0, q(t) F[x(t), v^*(t), t] = 0,$$

$$r(t) \geq 0, r(t) x(t) = 0, r(t) T[x(t), v^*(t), t] = 0;$$

$$(h) \quad F(T) \geq 0, p(T) x(T) = 0.$$

So far the propositions stated have been necessary conditions for the optimality of a policy. The situation is precisely analogous to the usual problem in calculus; the condition that a derivative be zero is necessary for a maximum but certainly not sufficient in general. However, the condition is sufficient if the function being maximized is concave. A basic property of concave functions is the following:

- (11) If  $f(x)$  is a concave function, then for any given point  $x^*$  and any other point  $x$  in the domain of definition,  $f(x) \leq f(x^*) + f_x^*(x - x^*)$ , where  $f_x^*$  is the row vector with components  $\partial f / \partial x_i$  evaluated at  $x^*$ .

Define the function

$$(12) \quad H^0(x, p, t) = \max_v H(x, v, p, t), \text{ where } v \text{ is constrained as in any of the Propositions 1-4.}$$

Then the concavity of  $H^0$  as a function of  $x$ , for given  $p$  and  $t$ , implies that the Pontryagin conditions are sufficient for optimality.

(This is a minor variation of a theorem of Mangasarian [1966].)

**Proposition 5.** If  $H^0$ , as defined in (12), is concave in  $x$  for given  $p$  and  $t$ , then any choice of instruments satisfying the conditions of any of Propositions 1-4 is optimal for the corresponding problem.

## Lecture 2

### Optimization with Infinite Horizon

For many purposes it is more convenient to introduce the fiction that the horizon is infinite. Certainly processes of capital accumulation for the economy as a whole have no natural stopping place in the definable future. At any given future date the state of the system (its capital structure) will have implications for the further future. If we choose to stop our analysis at any fixed date, it will be necessary, as already noted, to include in our utility functional some scrap value for the stock of capital at the end of the period. But the only logically consistent way of doing so is to determine the maximum utility attainable in the further future starting with any given stock of capital. Of course, the astronomers assure us that the world as we know it will come to an end in some few billions of years. But as elsewhere in mathematical approximations to the real world, it is frequently more convenient and more revealing to proceed to the limit to make a mathematical infinity in the model correspond to the vast futurity of the real world.

Formally, the only change in the statement of the model is to let  $T = +\infty$ . But going to the limit, here as elsewhere, involves some risks. The utility functional, now an improper integral, might not converge at all; and even if it does, there might not exist an optimal policy. However, it is still possible to state necessary conditions and sufficient conditions for optimality, though existence of an optimal policy may be difficult to guarantee, and also it is not yet known how to state the appropriate transversality conditions. An extensive discussion of a case of non-existence of an optimal path is given by Koopmans [1965, pp. 251-3].

If an optimal policy exists, then it can be shown that the arguments for the necessity conditions of Propositions 1-4, except for the transversality conditions, are still valid. In the cases of interest in economics, the transversality conditions (Propositions 1(c), 2(b), 3(g), or 4(g)) are in fact valid, but so far it is necessary to verify this in each case. The infinite-horizon statement of the transversality conditions of Propositions 2-4 is:

$$(13) \quad \lim_{t \rightarrow +\infty} p(t) \geq 0, \quad \lim_{t \rightarrow +\infty} p(t) \dot{x}(t) = 0.$$

The sufficiency theorems remain completely valid, with the transversality condition (13).

It is customary and reasonable to assume that future felicities are discounted; i.e., the felicity obtained at time  $t$  is multiplied by a discount factor  $\alpha(t)$ , which is ordinarily taken to be a decreasing function of  $t$ . This corresponds to the intuitive idea that future pleasures are counted for less today. The utility functional is rewritten:

$$(14) \quad \int_0^{+\infty} \alpha(t) U[x(t), v(t), t] dt.$$

Ordinarily, it is assumed that if the chosen policy leads to a constant felicity, (14) will converge. This is equivalent to the condition,

$$(15) \quad \int_0^{+\infty} \alpha(t) dt \text{ converges.}$$

If we follow the earlier line of argument we would be interested in the maximum utility obtainable starting at some time  $t_0$ , analogous to (4):

$$(16) \quad V(x, t_0) = \max \int_{t_0}^{+\infty} \alpha(t) U[x(t), v(t), t] dt, \text{ where } x(t_0) = x.$$

However, this means that felicities for times beyond  $t_0$  are being discounted to time 0. It is more natural to discount them to time  $t_0$ . Since one unit of felicity at time  $t_0$  is equivalent to  $\alpha(t_0)$  units at time 0, it is necessary to divide  $V(x, t_0)$  by  $\alpha(t_0)$  to obtain the current-value return function,

$$(17) \quad W(x, t_0) = V(x, t_0) / \alpha(t_0).$$

Previously we obtained the auxiliary variables,  $p(t)$ , as the marginal contributions of the state variables to the utility functional,  $p_1 = \partial V / \partial x_1$ . In the present context it seems more reasonable to define

$$(18) \quad p_1 = \partial W / \partial x_1 = (\partial V / \partial x_1) / \alpha.$$

In applying Proposition 4 (apart from the transversality condition) to the discounted infinite-horizon case, it is then necessary to replace  $U(x, v, t)$  by  $\alpha(t) U(x, v, t)$  and  $p(t)$  by  $\alpha(t) p(t)$ . The Hamiltonian becomes

$$(19) \quad \alpha(t) U(x, v, t) + \alpha(t) p(t) T(x, v, t) = \alpha(t) H(x, v, p, t),$$

where we now define the current-value Hamiltonian:

$$(20) \quad H(x, v, p, t) = U(x, v, t) + pT(x, v, t).$$

Then  $\alpha(t) H$  must replace  $H$  throughout the restatement of Proposition 4.

Since  $\alpha(t) > 0$ , the choice of instruments to maximize  $\alpha(t) H$  is the same as that to maximize  $H$ , so that Proposition 4(d) remains unchanged. If we interpret the Lagrange multipliers  $q$  and  $r$  as referring to the maximization of  $H$  as now defined subject to the constraints, then  $L$  must be replaced by  $\alpha(t) L$ . Proposition 4(e) becomes

$$\frac{d[\alpha(t)p_1]}{dt} = - \frac{\partial[\alpha(t)L]}{\partial x_1},$$

or

$$\dot{\alpha} p_1 + \alpha \dot{p}_1 = -\alpha (\partial L / \partial x_1) .$$

Divide through by  $\alpha$ , and define

$$(19) \quad \rho(t) = -\dot{\alpha}(t)/\alpha(t) .$$

Then

$$(20) \quad \dot{p}_1 = \rho(t) p_1 - (\partial L / \partial x_1) .$$

In economic terms,  $\rho(t)$  is a short-term interest rate, corresponding to the system of discount factors  $\alpha(t)$ . The definition (19) can be integrated back to yield the familiar form:

$$(21) \quad \alpha(t) = e^{-\int_0^t \rho(u) du} ,$$

if we adopt the convention that  $\alpha(0) = 1$ . If (20) is written

$$\dot{p}_1 + (\partial L / \partial x_1) = \rho(t) p_1(t) ,$$

it is the familiar equilibrium relation for investment in capital goods; the sum of capital gains and marginal productivity should equal the interest on the investment.

The infinite-horizon analogue of Proposition 4 (apart from transversality conditions) becomes:

Proposition 6. Let  $v^*(t)$  be a choice of instruments ( $t \geq 0$ ) which  $\int_0^{+\infty} \alpha(t) U[x(t), v(t), t] dt$  subject to the conditions (a), (b), and (c) of Proposition 4. If the Constraint Qualification holds, then there exist auxiliary variables  $p(t)$  satisfying (d) of Proposition 4;

$$(e) \quad \dot{p}_1 = \rho p_1 - (\partial L / \partial x_1), \text{ evaluated at } x = x(t), v = v^*(t), p = p(t), q = q(t), r = r(t).$$

where  $\rho(t) = -\dot{\alpha}(t)/\alpha(t)$ , and (f) and (g) of Proposition 4 hold

The sufficiency theorem, Proposition 5, remains valid if the transversality condition is replaced by (13) where, however,  $p(t)$  is replaced by  $\alpha(t) p(t)$ .

Proposition 7. In the notation of Propositions 4 and 6, if

$$H^0(x, p, t) = \max_v H(x, v, p, t),$$

where the maximization is over the range specified in Proposition 4(d), is a concave function of  $x$  for given  $p$  and  $t$ , then any policy satisfying the conditions of Proposition 6 and the transversality conditions,

$$\lim_{t \rightarrow +\infty} \alpha(t) p(t) \geq 0, \quad \lim_{t \rightarrow +\infty} \alpha(t) p(t) x(t) = 0,$$

is optimal.

It is frequently appropriate to make an assumption that the basic conditions of the optimization problem are stationary; the sequence of conditions to be encountered in the future is much the same as today or can be made so after some simple renormalizations. This property will be heavily exploited in our subsequent discussions. The basic stationarity assumptions are that the functions  $U(x, v, t)$ ,  $T(x, v, t)$ ,  $F(x, v, t)$ , and  $p(t)$  are all independent of time. With  $\rho$  constant, it follows from (21) that

$$(22) \quad \alpha(t) = e^{-\rho t},$$

and the convergence condition (15) becomes

$$(23) \quad \rho > 0.$$

Under the stationarity assumption, the current-value return function,  $W(x, t_0)$ , defined by (17), is in fact independent of  $t_0$ ; this can be seen by writing, in view of the previous remarks,

$$\begin{aligned}
W(x, t_0) &= (1/e^{-\rho t_0}) \max \int_{t_0}^{+\infty} e^{-\rho t} U[x(t), v(t)] dt \\
&= \max \int_{t_0}^{+\infty} e^{-\rho(t-t_0)} U[x(t), v(t)] dt.
\end{aligned}$$

Since the constraints  $F(x, v) \geq 0$  and the transition relations,  $\dot{x} = T(x, v)$ , do not involve time explicitly, it is clear that replacing  $t_0$  by 0, say, leaves completely unaffected the form of the optimal policy. This is an illustration of Bellman's [1957] well-known "principle of optimality."

But if  $W(x, t_0) = W(x)$ , independent of  $t$ , then from (18)  $p$  is completely determined by the state  $x$  in the following sense: Suppose we have two optimization problems of the type dealt with in Proposition 6 (but also satisfying the stationarity assumptions) which are identical in all respects except for initial conditions. Let  $x^1(t)$  and  $x^2(t)$  be the paths of the state variables along the optimal solutions for the two problems, respectively, and let  $p^1(t)$  and  $p^2(t)$  be the corresponding paths of the auxiliary variables. Then if  $x^1(t) = x^2(t')$ ,  $p^1(t) = p^2(t')$ .

Note that since  $p$  is determined by  $x$ , and  $U$  and  $T$  do not depend on  $t$ , along the optimal path  $H(x, v, p, t)$  is a function of  $x$  and  $v$  alone, and therefore the value of  $v$  which maximizes  $H$  depends only on  $x$ . The optimum policy can be represented as a strategy or feedback control, with  $v$  a function of  $x$ .

Also note that, for given  $x$ ,  $v$ , and  $p$ ,  $H$  is independent of  $t$ , and therefore Propositions 4 and 6(c), by itself implies that  $v^*$  is determined by  $x$  and  $p$ , independent of  $t$ . The stationarity assumptions then imply that  $t$  does not enter explicitly into the system of differential equations defined by (a) and (e). Such a system is termed autonomous.

Proposition 8. Under the assumptions and in the notation of Proposition 6, suppose in addition that

- (a)  $U(x,v,t) = U(x,v)$ ,  $T(x,v,t) = T(x,v)$ ,  $F(x,v,t) = F(x,v)$ , and  $p(t) = p$ , all independent of  $t$ .

Then

- (b) the optimal policy,  $v^* = v^*(x)$ , and the values of the auxiliary variables,  $p$ , along the optimal path, are functions of  $x$  alone, independent of  $t$  for given  $x$ ;
- (c) the system of differential equations defined by (a), (d), and (e) is autonomous.

For an autonomous system, considerable interest usually relates to its stationary point or equilibrium, where all motion ceases, i.e., the values of  $x$  and  $p$  for which  $\dot{x} = 0$  and  $\dot{p} = 0$ . This notion in economics is that of long-run stationary equilibrium (as opposed to temporary or short-run equilibrium in which capital stocks are given). In the present system an equilibrium is defined by  $x^*$ ,  $p^*$ ,  $v^*$  satisfying the conditions:

$$T(x^*, v^*) = 0,$$

$$p p_1^* = L_{x_1}^*,$$

$$v^* \text{ maximizes } H(x^*, v, p^*) \text{ under the constraints } F(x^*, v) \geq 0,$$

$$T_1(x^*, v) \geq 0 \text{ if } x_1^* = 0.$$

If the initial state of the system is  $x^*$ , then all the conditions of Proposition 6 can be satisfied by setting  $x(t) = x^*$ ,  $v(t) = v^*$ ,  $p(t) = p^*$  for all  $t$ . It may be asked under what conditions this solution is optimal. More generally, suppose we can find a path satisfying the conditions of Proposition 1 which converges to the stationary values; when is

this optimal?

For simplicity of reference, define a Pontryagin path as a system,  $x(t)$ ,  $p(t)$ ,  $v^*(t)$ , satisfying the conditions of Proposition 6.

Proposition 9. Let  $x(t)$ ,  $p(t)$ ,  $v^*(t)$  be a Pontryagin path for the problem of Proposition 6. Suppose further that the concavity hypothesis of Proposition 7 and the stationarity hypothesis of Proposition 8, with  $\rho > 0$ , are satisfied. Then, if  $x(t)$  and  $p(t)$  converge to an equilibrium,  $x^*$ ,  $p^*$ , where  $p^* \geq 0$ , they constitute an optimal path.

Proof: From Proposition 7 it suffices to note that the transversality condition  $\alpha(t) p(t) x(t) \rightarrow 0$  is satisfied. But  $p(t)$  and  $x(t)$  converge to finite limits, and  $\alpha(t) = e^{-\rho t}$  approaches zero since  $\rho > 0$ .

It should be remarked, however, that (a) there may be more than one equilibrium, and (b) there may exist optimal paths which do not converge to any finite equilibrium; for examples, see Kurz [1965 and 1967, Section B].

### Lecture 3

#### Optimal Investment Planning in a One-Commodity Model

In this lecture we review in detail the simplest possible capital accumulation model, first studied by Ramsey [1928]; for further important contributions, see Mirrlees [1967] and Koopmans [1965]. We assume there is only one commodity, which can be either consumer or invested. We take the viewpoint of a government which is in a position to control the economy completely and to plan perfectly so as to optimize with respect to all possible instruments of the economic system - in this case, only investment and consumption (which are subject to the constraint that their sum not exceed total output).

We first assume a constant population and a constant labor force. The felicity at any moment is taken to be a function of consumption,  $C$ , only. Then the aim of the economy is to maximize

$$(24) \quad V = \int_0^{+\infty} e^{-\rho t} U[C(t)] dt,$$

where  $\rho$  is the rate of interest on felicity,  $C(t)$  is consumption at time  $t$ , and  $U(C)$  is the felicity derived from consumption  $C$ . It is assumed that

$$(25) \quad U(C) \text{ is strictly concave and increasing.}$$

The output at any moment of time is a function of the stock of capital and of the labor force. Since the latter is assumed constant, we assume simply

$$(26) \quad Y(t) = F[K(t)],$$

where  $Y(t)$  is output at time  $t$  and  $K$  is the stock of capital. With the labor force held constant, increases in capital may be supposed to yield lower and lower returns; also it is assumed that capital is indispensable to production.

$$(27) \quad F \text{ is strictly concave, } F(0) = 0.$$

It is not necessarily assumed that  $F(K)$  is increasing; for example, if the stock of capital depreciates at a rate proportional to its quantity, then the depreciation ought properly to be subtracted from the gross output to get a true measure of net output available for consumption and net investment (increase of the capital stock). It is possible that if  $K$  is very large, the marginal gross output of an additional unit may be less than the depreciation on that unit.

Finally, the accumulation of capital is precisely investment,  $I$ ,

$$(28) \quad \dot{K} = I,$$

and the conservation of product flow implies that consumption and investment, together, do not exceed output, i.e.,  $C + I \leq Y$  or, in view of (26),

$$(29) \quad F(K) - C - I \geq 0.$$

It also follows from the very definition of capital that it cannot be negative;  $K(t) \geq 0$ , all  $t$ .

At present it will be assumed that  $I$  may be positive, zero, or negative; the latter means that existing capital can be turned into consumption goods. The case where  $I$  is necessarily non-negative will be considered later. It will be assumed that  $C \geq 0$ ; but to simplify matters it will be also assumed for the moment that

$$(30) \quad U'(0) = +\infty,$$

which, as will be seen, implies that the choice of the instrument  $C$  at any moment of time will necessarily be positive, so that the non-negativity constraint is ineffective.

Propositions 6-9 can be applied to this model. The state of the system is represented by the single variable,  $K$ . There are two instruments,  $C$  and  $I$ . The felicity function depends only on the one instrument  $C$ ; the transition function (28) depends only on  $I$ . The choice of  $C$  and  $I$  is constrained by (29), which corresponds to Proposition 6(b) or 4(b). There will be one auxiliary variable,  $p$ , so that the current-value Hamiltonian is

$$H = U(C) + pI,$$

and the Lagrangian,  $L$ , is

$$\begin{aligned} (31) \quad L &= U(c) + pI + q[F(K) - C - I] \\ &= [U(C) - qC] + (p-q)I + qF(K). \end{aligned}$$

By Proposition 6(g) or 4(g),  $C$  and  $I$  must be chosen so that  $\partial L / \partial C = 0$  and  $\partial L / \partial I = 0$ . The latter implies that

$$(32) \quad p = q.$$

The former implies that  $U'(C) = q$  and, by (32),

$$(33) \quad U'(C) = p.$$

Because of (30) and the concavity of  $U(C)$ , it is assured that the solution to (33) will be positive.

The auxiliary equation, Proposition 6(e), becomes

$$\begin{aligned} \dot{p} &= \rho p - q(\partial L / \partial K), \\ (34) \quad \dot{p} &= [\rho - F'(K)]p, \end{aligned}$$

in view of (31) and (32). Since  $U' > 0$  by (25),  $p > 0$  by (33), and

$q > 0$  by (32); the constraint (29) is effective, so that from (28)

$$(35) \quad \dot{K} = F(K) - C(p),$$

where

$$(36) \quad C(p) \text{ is the solution of (33).}$$

Equations (34) and (35) are a pair of autonomous equations. An equilibrium is defined by  $\dot{p} = 0$  and  $\dot{K} = 0$ . But  $\dot{p} = 0$  means

$$p = 0 \text{ or } F'(K) = p.$$

Since  $U' > 0$ , the first alternative is impossible from (33). Let  $K^\infty$ ,  $p^\infty$  be the equilibrium values of  $K$  and  $p$ , respectively, and let  $C^\infty$  and  $I^\infty$  be the values of the instruments at the equilibrium. It has just been shown that

$$(37) \quad F'(K^\infty) = p.$$

Since  $F$  is strictly concave,  $F'$  is strictly decreasing. It will be assumed that (37) has a solution with  $K^* > 0$ . This is equivalent to assuming that  $F'(0) > p$ ,  $F'(K) < p$  for  $K$  sufficiently large.

From (28) and the definition of equilibrium,

$$(38) \quad I^\infty = 0.$$

From (35) and (36), with  $\dot{K} = 0$ ,

$$(39) \quad C^\infty = F(K^\infty),$$

$$(40) \quad p^\infty = U'(C^\infty).$$

Consider now all solutions of the differential equations (34) and (35). Their movements may be represented in a phase diagram (Figure 1). Since  $F'(K)$  is decreasing,  $p - F'(K)$  is increasing; from (34), then,  $\dot{p} > 0$  if  $K > K^\infty$ ,  $\dot{p} < 0$  if  $K < K^\infty$ . Since  $U'$  is a decreasing function,

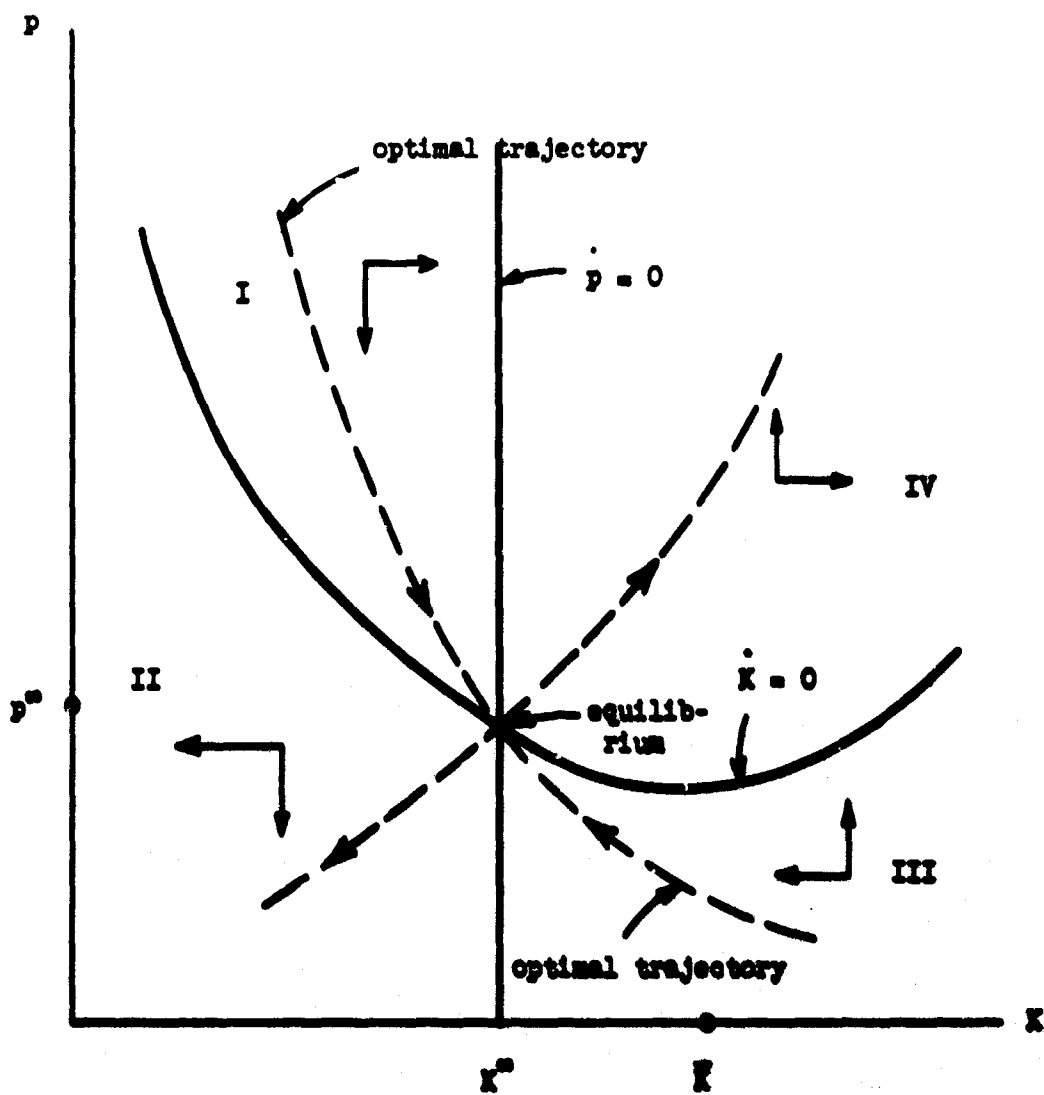


Figure 1

It follows from (36) and (33) that  $C(p)$  is a decreasing function of  $p$ . The curve for which  $\dot{K} = 0$  is, from (35), defined by the equation

$$F(K) = C(p).$$

This equation can be solved uniquely for  $p$  in terms of  $K$ ; call the solution  $\pi(K)$ . Since  $F(K)$  is concave, and  $F'(K^\infty) = \rho > 0$ ,  $F(K)$  is either always increasing or increasing up to a value  $\bar{K} > K$ ; therefore,  $\pi(K)$  is decreasing for all  $K$  or else decreasing to  $\bar{K}$  and increasing thereafter. Further, for fixed  $K$ ,  $\dot{K} = F(K) - C(p)$  is an increasing function of  $p$  so that  $\dot{K} > 0$  above the curve and  $\dot{K} < 0$  below. The components of the directions of movement in the four quadrants into which the diagram is divided by the curve  $\dot{K} = 0$  and the vertical line  $\dot{p} = 0$  are indicated in Figure 1 by arrows; note that  $\dot{K} > 0$  is a movement to the right, and  $\dot{p} > 0$  an upward movement.

Regions II and IV are traps in the sense that a Pontryagin path which enters either of these regions never leaves it. Further, any path which comes to a boundary of either region must enter that region and then remain in it permanently. It will now be shown that a path which enters either region cannot be optimal.

Consider first region IV. Without loss of generality, suppose that the path is in region IV at time 0. Then  $p > 0$ ,  $\dot{K} > 0$ , all  $t$ . Since  $K(0) > K^\infty$ ,  $K(t) \geq K(0) > K^\infty$ , all  $t$ . Since  $F'$  is decreasing,  $F'[K(t)] \leq F'[K(0)] < F'(K^\infty) = \rho$ , so that, from (34),  $\dot{p}/p = \rho - F'[K(t)] \geq \epsilon = \rho - F'[K(0)] > 0$ ;  $p(t) \geq p(0) e^{\epsilon t}$  by integration, so that, certainly,

$$p(t) > p^\infty \text{ for all } t \geq t_0.$$

for some  $t_0$ . Since  $C(t) = C(p(t))$  is a decreasing function of  $p$ ,

$$C(t) < C(p^{\infty}) = C^{\infty} \text{ for } t \geq t_0.$$

Since  $K(t_0) > K^{\infty}$ , it follows that we can always improve on the given path by consuming the capital stock (disinvesting) in some interval beginning at  $t_0$  until  $K$  diminishes to  $K^{\infty}$ , after which the equilibrium policy,  $C = C^{\infty} = F(K^{\infty})$ ,  $K = K^{\infty}$ , is maintained perpetually.

Now consider any trajectory in region II. By the same reasoning  $\dot{p}/p \leq \epsilon$ , where now  $\epsilon = \rho - F'[K(0)] < 0$ . But then  $p(t) \rightarrow 0$ , which implies  $C[p(t)] \rightarrow +\infty$ . Since  $F(K)$  is uniformly bounded on the closed interval  $(0, K^{\infty})$ ,  $K = F[K(t)] - C[p(t)] \leq \delta < 0$  from some time on. Then  $K(t)$  must become zero at some finite time. Since  $C > 0$ , then,  $I = F(0) - C < 0$ , and  $K$  will become negative, violating the non-negativity of  $K$ .

Consider now a path starting in quadrant I. If it stays in quadrant I forever, then both  $p$  and  $K$  are bounded from below. Since both are decreasing, they approach limits which, by a general theorem on differential equations, can only be the equilibrium values. By Proposition 9 such a path is necessarily optimal. If the path did not remain in quadrant I for all  $t$ , then it reaches either the boundary with quadrant IV or that with quadrant II; then, as already noted, the path cannot be optimal.

Similarly, any path in quadrant III which remains there forever approaches the equilibrium and is optimal; any other path is non-optimal.

It only remains to argue that, for any initial  $K = K(0)$ , there is a corresponding  $p(0)$ , with the point  $(K, p(0))$  in quadrant I or quadrant III according as  $K < K^{\infty}$  or  $K > K^{\infty}$ , such that the Pontryagin path starting at that point approaches the equilibrium. Such a path is certainly optimal.

The approach makes use of the fact that, under the stationarity assumptions of this problem,  $p$  and the optimal instruments  $C$  and  $I$  are functions of  $K$ . Divide (34) by (35) to see that

$$(41) \quad dp/dK = p[p - F'(K)]/[F(K) - C(p)].$$

Consider, for values of  $K \leq K^*$ , solutions of (41) which intersect the line  $K = K^*$  above the equilibrium, i.e., for  $p(K^*)$  a prescribed value greater than  $p^*$ . Such a solution can be continued for smaller and smaller values of  $K$ . We first note that it can never cross the curve  $\dot{K} = 0$ , which has been written  $p = \pi(K)$ . Let  $p(K)$  be the solution of (41), and suppose it intersected the curve  $p = \pi(K)$  at  $\tilde{K} < K^*$ . Then  $p'(K) \rightarrow -\infty$  as  $K \rightarrow \tilde{K} + 0$ . But  $p(K) > \pi(K)$  for  $K$  in a right-hand neighborhood of  $\tilde{K}$ , and therefore  $p'(\tilde{K}) \geq \pi'(\tilde{K})$ , a contradiction since  $\pi'(\tilde{K})$  is certainly finite.

Since the denominator of (41) is finite and the numerator is bounded from above, it is clear that the solution of (41) can be continued for all positive values of  $K \leq K^*$ . There is one such solution for each value of  $p(K^*) > p^*$ . These solutions never cross because of the uniqueness of solutions of this differential equation away from the equilibrium point. Hence, for any given  $K < K^*$ , there is a lower bound,  $p(K)$ , on the values of  $p$  for which there exists a solution of (41) passing through  $(K, p)$  and for which  $p(K^*) > p^*$ . It is obvious and can easily be demonstrated that  $p(K)$  also satisfies (41), and that  $p(K^*) = p^*$ . This path in  $(p, K)$ -space defines the optimal trajectory. If  $K(0) < K^*$ , choose  $p(0) = p[K(0)]$ . Then the points of the time solution,  $p(t)$ ,  $K(t)$ , for (34) and (35) move along the trajectory  $p(K)$  and converge to the equilibrium.

The solution in this form is very convenient, for the choice of the instruments,  $C$  and  $I$ , is determined as a function of  $K$  by (33) and (29) (with equality).

The analysis in quadrant III is the same, except that we find for each  $K > K^m$  the upper bound of  $p$ -values for which the solution of (41) passes below the equilibrium.

It should, however, be noted that we could apply the same procedure in quadrants II and IV; but then the limiting solution would be the divergent dashed curves marked in Figure 1.

The optimal solution, then, is defined by a solution of (41) which passes through the equilibrium; but there are two such solutions. The equilibrium is a singular point of (41), so the solution through that point need not be unique. In this case it is clear that the optimal solution is identified as the one with the negative slope at the equilibrium.

We will analyze the non-uniqueness at equilibrium a little more closely. The right-hand side of the differential equation (41) is, strictly speaking, not defined at  $K = K^m$ ,  $p = p^m$ , since both numerator and denominator vanish. Since  $p$  is to be a function of  $K$ , both numerator and denominator are functions of  $K$ , directly and through  $p$ .

Let

$$(42) \quad \phi(K) = p[p - F'(K)], \quad \psi(K) = F(K) - C(p),$$

$$(43) \quad p'(K) = \phi(K)/\psi(K).$$

Since both  $\phi$  and  $\psi$  vanish at  $K = K^m$ , we can define  $p'(K)(=dp/dK)$  there by L'Hôpital's rule:

$$(44) \quad p'(K^m) = \phi(K^m)/\psi'(K^m),$$

and it remains to evaluate these derivatives.

$$\phi'(K) = p[-F''(K)] + [\rho - F'(K)] p'(K).$$

From (37),

$$\phi'(K^\infty) = -p^\infty F''(K^\infty).$$

$$\psi'(K) = F(K) - C'(p) p'(K).$$

Again make use of (37).

$$\psi'(K^\infty) = \rho - C'(p^\infty) p'(K^\infty).$$

Substitute into (44).

$$p'(K^\infty) = \frac{-p^\infty F''(K^\infty)}{\rho - C'(p^\infty) p'(K^\infty)},$$

or, clearing fractions,

$$(45) \quad -C'(p^\infty) [p'(K^\infty)]^2 + \rho p'(K^\infty) + p^\infty F''(K^\infty) = 0,$$

a quadratic equation in the slope of the solution to (41) which passes through the equilibrium. Since  $C(p)$  is decreasing, the coefficient of the quadratic term is positive. Since  $F'' < 0$ , the constant term is negative. Thus, the product of the roots is negative, which implies that both are real, with one positive and one negative. As already noted, the negative root is the appropriate one.

Since  $C(p)$  is defined by (36) and (33), we must have  $U''[C(p)] C'(p) = 1$ , so that  $C'(p^\infty) = 1/U''(C^\infty)$ .

Proposition 10. Suppose the aim of the economic system is to maximize  $\int_0^\infty e^{-\rho t} U[C(t)] dt$ , where  $\rho > 0$ , subject to the conditions  $\dot{K} = I, C + I \leq F(K), K \geq 0$ , where  $U(C)$  is a strictly concave increasing function and  $F(K)$  is a strictly concave function with  $F(0) = 0$ , and

also assume  $K(0)$  given. Define  $K^\infty, C^\infty, p^\infty$  by the relations  $F'(K^\infty) = p, C^\infty = F(K^\infty), p^\infty = U'(C^\infty)$ .

Then the optimal strategy can be characterized by finding that solution  $p(K)$  of the differential equation (41) for which  $p(K^\infty) = p^\infty$  and for which  $p'(K^\infty)$  is the negative root of the quadratic equation  $[-1/U''(C^\infty)] [p'(K^\infty)]^2 + p p'(K^\infty) + p^\infty F''(K^\infty) = 0$ . Then, for any  $K$ ,  $C$  is so chosen that  $U'(C) = p(K)$ , and  $I = F(K) - C$ .

Proposition 10 has been stated without the hypothesis,  $U'(0) = +\infty$ , which was used in the proof. It will be an interesting exercise in the use of Proposition 6 to consider the case where  $U'(0)$  is finite. In this case the constraints  $C \geq 0$  and  $h \geq 0$  may become effective. Consider the first for regions in which  $K > 0$ . Let  $w$  be the multiplier associated with the constraint,  $C \geq 0$ . Then (31) is modified to read

$$(46) \quad \begin{aligned} L &= U(C) + pI + q[F(K) - C - I] + wC \\ &= [U(C) - (q-w)C] + (p-q)I + qF(K), \end{aligned}$$

where

$$(47) \quad w \geq 0, wC = 0.$$

If  $C = 0$ , then the condition  $\partial L / \partial C = 0$  becomes

$$U'(0) = U'(C) = q - w \leq q.$$

We still have the condition  $p = q$ , so that  $C = 0$  if  $p \geq U'(0)$ . The system (34) and (35) is still valid, but the definition of  $C(p)$  is slightly modified;

$$\begin{aligned} C(p) &\text{ is the solution of the equation, } U'(C) = p \\ &\text{ if } p \leq U'(0), \\ C(p) &= 0 \text{ if } p > U'(0). \end{aligned}$$

The previous analysis is completely unchanged; in Figure 1, the curve  $\dot{K} = 0$  intersects the p-axis at  $p = U'(0)$  instead of being asymptotic to it. Since the optimal trajectory lies above the curve  $\dot{K} = 0$ , there will be a  $\underline{k} > 0$  for which  $p(\underline{k}) = U'(0)$ . For  $k \leq \underline{k}$ ,  $C$  will then be zero.

Now consider the possibility that the constraint  $K = 0$  becomes effective. As has already been seen, this question arises only for paths which start in or have entered region II. It will be shown, even with  $U'(0)$  finite, such paths are non-optimal.

Recall the basic definition (18) of  $p$  as  $dW/dK$ , where  $W$  is maximum value of the utility functional if the initial state is  $K$ . Clearly, in this model an increase in  $K$  is always beneficial; given an increase in  $K$ , one can always consume a somewhat higher amount for some period until the value of  $K(t)$  falls to that on the original path, and then follow the latter thereafter. Hence,  $p$  must be positive.

Now consider a path that has reached the p-axis at time  $t_0$  from region II. Since the initial value of  $p$  was finite, the time to reach the p-axis was finite, and the right-hand side of (34) is bounded over this path,  $p(t_0)$  is finite. Now the constraint  $K \geq 0$  becomes effective and therefore the constraint  $I \geq 0$  is imposed. The constraints,

$$C \geq 0, I \geq 0, C + I \leq F(K) = F(0) = 0,$$

imply that  $C$  and  $I$  are 0; from the latter, it follows that  $K(t) = 0$  for all  $t \geq t_0$ . The Lagrangian (46) is modified by the addition of a term corresponding to the constraint  $I \geq 0$ , with multiplier  $r$ .

$$\begin{aligned} \mathcal{L} &= U(C) + pI + q[F(K) - C - I] + wC + rI \\ &= [U(C) - (q-w)C] + (p+r-q)I + qF(K), \end{aligned}$$

where  $r \geq 0$  when  $I = 0$ . No longer does the equality  $p = q$  hold; instead,

$$p = q - r \leq q.$$

Equation (34) is modified to

$$\dot{p} = \rho p - qF'(K) = \rho p - qF'(0) \quad \text{for } t \geq t_0.$$

Recall that  $F'(0) > \rho$ . Use in turn the inequalities  $p \leq q$ ,  $U'(0) \leq q$ .

$$\dot{p} = \rho p - qF'(0) \leq q[\rho - F'(0)] \leq U'(0) [\rho - F'(0)].$$

The last term is a negative constant. Hence,  $p$  must become negative in finite time, which is a contradiction to the assumption of the optimality of the path being studied.

It may be asked what happens if the initial stock of capital is 0. The only feasible path is that of zero investment and consumption. The argument just given would show that for any finite  $p(0)$ ,  $p(t)$  would become negative eventually. The answer evidently is that  $p(0)$  must be chosen  $+\infty$  initially, and then  $p(t)$  would remain  $+\infty$ .

#### Lecture 4

##### Further Aspects of the Ramsey Problem: Irreversibility; Growth of Population and Labor Force and Technical Change

It is sometimes reasonable to argue that investments, once made in physical form, cannot be converted into consumer goods. Hence, investment should be irreversible, i.e., subject to the constraint  $I \geq 0$ . On the other hand, there is in real life a somewhat more subtle way in which capital can, within limits, be run down and permit more consumption; namely, capital goods depreciate and failure to replace them constitutes a way of increasing consumption at the expense of capital.

A reasonable assumption about depreciation is that a fixed fraction of the existing capital becomes useless in each time period. Thus, the net rate of increase of capital is the amount of (gross) investment, i.e., new output devoted to capital uses, less the amount of depreciation. This amounts to replacing (28) by

$$(48) \quad \dot{K} = I - \delta K,$$

for some  $\delta > 0$ . We also assume that investment is non-negative,

$$(49) \quad I \geq 0.$$

Otherwise the model is identical with that of the last lecture, including (24) and (29), with the assumptions (25), (27), and (30) (though the last is dispensable). Let  $p$  be the auxiliary variable corresponding to (48),  $q$  the multiplier corresponding to (29), and  $s$  that corresponding to (49). The Lagrangian becomes

$$(50) \quad \begin{aligned} L &= U(C) + p(I - \delta K) + q[F(K) - C - I] + sI \\ &= [U(C) - qC] + (p+s-q)I + qF(K) - p\delta K, \end{aligned}$$

where

$$(51) \quad s \geq 0, \quad sI = 0.$$

Equating derivatives with respect to  $C$  and  $K$  to zero yields

$$(52) \quad U'(C) = q,$$

$$q = p + s,$$

which can be combined with (51) to yield

$$(53) \quad p \leq q; \quad \text{if } p < q, \quad \text{then } I = 0.$$

The auxiliary equation is

$$(54) \quad \dot{p} = (\rho + \delta) p - qF'(K).$$

Since the problem is stationary, we know that the instruments and the auxiliary variable are functions of  $K$ . From (53), for any given  $K$ , there are two possibilities: either  $p = q$ , or  $p < q$  with  $I = 0$ . Therefore, the  $K$ -axis is divided, in general, into alternating blocked and free intervals:

$$(55) \quad I = 0, \quad C = F(K), \quad q = U'[F(K)] > p \quad \text{on a blocked interval};$$

$$(56) \quad I \geq 0, \quad C \leq F(K), \quad q = p \geq U'[F(K)] \quad \text{on a free interval}.$$

From (48),  $\dot{K} = -\delta K < 0$  if  $I = 0$ . Hence, the system cannot have an equilibrium in a blocked interval. Then  $p = q$  at an equilibrium and, from (54), (48), and (29), the following relations hold at equilibrium:

$$(57) \quad \begin{aligned} F'(K^*) - \delta &= \rho, \\ I^* &= \delta K^*, \\ C^* &= F(K^*) - \delta K^*, \\ p^* &= U'(C^*). \end{aligned}$$

It will be observed that the equilibrium is the same as that for the reversible Ramsey problem, where  $F(K)$  is replaced by  $F(K) - \delta K$ . Indeed, if we define

$$I_N = I - \delta K,$$

then, if the constraint (49) is ineffective, the problem is identical with the reversible Ramsey problem, with  $I$  replaced by  $I_N$ . For the reversible Ramsey problem the optimal policy would have  $I_N \geq 0$  for  $K \leq K^*$ , and therefore  $I = I_N + \delta K \geq 0$ . Hence, starting with any such  $K$ , if the optimal policy for the reversible Ramsey problem is followed, it will always satisfy constraint (49) and therefore remain feasible under irreversibility. It will therefore remain optimal. Indeed, the same must be true for  $K$  in some right-hand neighborhood of  $K^*$ , for while  $I_N = 0$ ,  $I_N + \delta K > 0$  for  $K = K^*$  and, by continuity,  $I \geq 0$  for  $K^* \leq K \leq \underline{K}$  for some  $\underline{K}$  (which might even be  $+\infty$ ). On the optimal path  $K$  decreases in this interval, and therefore  $K$  never goes outside the interval, so that the optimal path for the reversible Ramsey path is still feasible and therefore optimal for  $K \leq \underline{K}$ .

The general method for finding the optimal strategy can now be sketched. As before, we are interested in the differential equation defining  $p'(K) = dp/dK$ . From (48) and (54),

$$(58) \quad p'(K) = [(\rho + \delta)p - qF'(K)] / (1 - \delta K).$$

Here  $q$  and  $I$  can be determined as functions of  $K$  and  $p$  from (55) and (56). In the neighborhood of the equilibrium, the solution, as noted, is the same as for the reversible case. From (55) and (56), (58) specializes in the two kinds of intervals as follows:

$$(59) \quad p'(K) = p[\rho + \delta - F'(K)]/[F(K) - \delta K - C(p)] \text{ in a free interval;}$$

$$(60) \quad p'(K) = [(\rho + \delta) p - U'[F(K)] F'(K)]/(-\delta K) \text{ in a blocked interval.}$$

Let

$$(61) \quad r = U'[F(K)]/p.$$

From (55) and (56),

$$(62) \quad r \leq 1 \text{ on a free interval, } r > 1 \text{ on a blocked interval.}$$

We know there is a free interval,  $(0, \underline{K})$ , with  $\underline{K} > K^*$ . We therefore solve (59) around the equilibrium and continue it first for all smaller values of  $K$ . Then continue it for larger values of  $K$  until  $r$  reaches 1. The  $K$ -value where this occurs is  $\underline{K}$ , and there is a calculated value of  $p$ ,  $p(\underline{K})$ . We then solve (60) with this starting point until  $r$  comes down to 1 from above. At this point we start a new free interval, and solve (59), but with the starting point being that achieved at the end of the previous blocked interval. This process can be continued indefinitely.

Thus, the problem is capable of numerically meaningful solution. Analytically sharper characterization cannot be obtained in general, though more specific hypotheses imply some limits on the numbers of blocked and free intervals. In particular, though, it can be shown that it is possible to have a denumerable or arbitrarily large finite number of alternations between free and blocked intervals. For these and other results, see Arrow and Kurz [1967].

Another modification of the Ramsey model consists of allowing for growth in population and labor force and for technological change. Under certain simple but by no means absurd assumptions, these factors

can be introduced into the Ramsey model by a simple reinterpretation of variables.

Population by itself affects the utility functional. Let  $N(t)$  be the number of individuals at time  $t$ . Assume for simplicity that the aggregate amount of consumption at any time  $t$  is divided equally among the existing population. Assume also that each individual has the same felicity function. Then the felicity of any individual at time  $t$  is  $U[C(t)/N(t)]$ . Since there are  $N(t)$  individuals, it is reasonable to conclude that the total felicity of society at time  $t$  is  $N(t) U[C(t)/N(t)]$ , and the utility functional then is

$$(63) \quad \int_0^{\infty} e^{-\rho t} N(t) U[C(t)/N(t)] dt.$$

The production possibilities of society are of course influenced by the size of the labor force. This effect has been ignored until now because the labor force has been assumed constant. The growth of the labor force is roughly proportional to that of population, but it will be convenient to ignore this relation for the moment. We assume in any case that the size of the labor force is a known function of time, independent of the instruments or the state variables. Let  $L(t)$  be the number of workers at time  $t$ . For any given supplies of capital and labor, output is determined by the production function

$$(64) \quad Y = F(K, L),$$

where it is assumed that

$$(65) \quad F \text{ is concave and homogeneous of degree 1, and } F(0, L) = 0.$$

The property of homogeneity of degree 1 is known to economists as constant returns to scale; if labor and capital are varied in the same

proportion, then the same productive methods can be employed, with only the scale changed, and therefore output can be changed in the same proportion. This assumption is not fully true but may be accepted as an approximation. The assumption  $F(0,L) = 0$  amounts to saying that capital is indispensable in production.

The transition equation of the system is still

$$(48) \quad \dot{K} = I - \delta K.$$

The constraint on the instruments,  $C$  and  $I$ , is now

$$(66) \quad F(K,L) \geq C + I.$$

Technological progress can be stated formally as

$$(67) \quad Y = F(K,L,t),$$

that is, the output obtainable from fixed amounts of capital and labor varies with time, presumably increasing. A particular hypothesis about technological progress for which there is some evidence is that it is labor-augmenting, which has the more specific form:

$$Y = F[K, A(t) L],$$

that is, each worker at time  $t$  can do, in every way, exactly what  $A(t)$  workers could do at time  $0$ . In this form, however, we can see that we may as well retain (64) where, however, it is understood that  $L$  now represents not the number of workers in the usual sense but the number of efficiency-equivalent workers. Thus, in the new definition,  $L$  can and usually will be increasing more rapidly than  $N$ .

The Lagrangian is

$$(68) \quad H = N(t) U[C(t)/N(t)] + p(I - \delta K) + q[F(K,L) - C - I].$$

The necessary conditions, with  $I$  unrestricted as to sign, become

$$(69) \quad p = q, U'[C(t)/N(t)] = p,$$

$$(70) \quad \dot{p} = p[\rho + \delta - (\partial F / \partial K)].$$

The system looks much as it did before, but it is not autonomous since time enters explicitly through  $N(t)$  and through  $L(t)$  in  $\partial F / \partial K$ . It is possible to use a non-autonomous system, but autonomous systems are much more convenient; with appropriate changes of variables, together with additional assumptions, it is possible to state the system in autonomous form.

Since labor is growing we can hardly expect an equilibrium in terms of the original variables, but it is reasonable to suppose there will be one in terms of ratios to the labor force. Divide all variables by  $L$ , and let small letters denote the resulting intensive magnitudes:

$$(71) \quad c = C/L, k = K/L, i = I/L.$$

Let

$$(72) \quad f(k) = F(k, 1).$$

Then, from (65),

$$(73) \quad f(k) \text{ is concave, } f(0) = 0,$$

$$(74) \quad F(K, L) = LF(K/L, 1) = Lf(K/L),$$

so that  $f(k)$  expresses the output per (effective) worker as a function of the capital per worker. Differentiate (74) partially with respect to  $K$ .

$$(75) \quad F_K = L(1/L) f'(K/L) = f'(K/L).$$

(70) can then be written

$$(76) \quad \dot{p}/p = \rho + \delta - f'(k).$$

Since  $\log k = \log K - \log L$  (natural logarithms),

$$\dot{k}/k = (\dot{K}/K) - (\dot{L}/L).$$

Multiply through by  $k$ , note that  $k/K = 1/L$ , substitute from (48), and use the definitions (71).

$$(77) \quad \dot{k} = i - (\gamma + \delta)k,$$

where

$$(78) \quad \gamma = \dot{L}/L,$$

frequently referred to as the natural rate of growth of the economy (remember that  $L$  has been so defined as to reflect technical progress as well as labor force growth). Divide through in (66) by  $L$ , and use the definition (71) and (72).

$$(79) \quad f(k) \geq c + i.$$

The equality will certainly always hold in (79). Elimination of  $i$  between (77) and (79) yields

$$(80) \quad \dot{k} = f(k) - (\gamma + \delta)k - c.$$

Now define

$$(81) \quad g(k) = f(k) - (\gamma + \delta)k;$$

from (73),  $g(k)$  is concave,  $g(0) = 0$ . Then (76) and (80) can be written:

$$(82) \quad \dot{p}/p = \rho - \gamma - g'(k),$$

$$(83) \quad \dot{k} = g(k) - c.$$

Finally, (69) can be written

$$(84) \quad U'[(L/N)c] = p.$$

The system (82-84) would be autonomous if the following two conditions are satisfied:

(85)  $\gamma$  constant,

(86)  $L(t)/N(t)$  constant.

This is the case of no technological progress and a constant rate of population and labor force growth. Then the equations have exactly the same form as those for the Ramsey model, with  $K$ ,  $C$ ,  $F(K)$ , and  $\rho$  replaced by  $k$ ,  $c$ ,  $g(k)$ , and  $\rho - \gamma$ , respectively. The importance of the last substitution must be stressed. The optimality analysis of the Ramsey case made use of the hypothesis,  $\rho > 0$  to show that the transversality conditions were satisfied. This condition seems reasonable. But in the case of growth, the corresponding condition is  $\rho > \gamma$ ; this is somewhat odd because the value of  $\rho$  is a value judgment while that of  $\gamma$  is an empirical fact. There seems no intrinsic reason why the inequality should hold in one direction or the other.

It must be remarked, moreover, that the hypothesis cannot be essentially weakened. In the Ramsey model without growth it can be shown that if  $\rho < 0$ , there is no optimal path in any meaningful sense; for a detailed analysis see Koopmans [1965, p. 251-2 and 279-85]; as just seen, the same result holds if the economy is growing at a constant rate  $\gamma$  and  $\rho < \gamma$ . The borderline case,  $\rho = 0$  in the model without growth or  $\rho = \gamma$  in a growing economy, has been studied in considerable detail by Ramsey [1928], Koopmans [1965, pp. 239-43 and 269-75], von Weizsäcker [1965]. Alternative definitions of optimality are possible since the utility functional need not converge, and in general the existence of an optimal program in the borderline case depends on the specific properties of the

production function.

To allow for technological progress, we wish to relax (86) and allow the ratio  $L(t)/N(t)$  to be increasing. We still wish to arrive at an autonomous system. In the system (82-84) it is (84) which will no longer be autonomous. In general there is no transformation of the variables in (82-84) which will make the system autonomous, but such a transformation is possible if  $U'$  is homogeneous of some degree. Note that  $U'$  must be decreasing; therefore it must be homogeneous of some negative degree, say  $-\sigma$ .

(87) Assume that  $U'(z)$  is homogeneous of degree  $-\sigma$ ,  $\sigma > 0$ .

We also assume, to replace (86),

(88)  $L(t)/N(t)$  has a constant rate of growth,  $\tau$ ,

which may be interpreted as the rate of (labor-augmenting) technological progress.

From (87) and (84),

$$[L(t)/N(t)]^{-\sigma} U'(z) = p,$$

or

$$U'(z) = p[L(t)/N(t)]^{\sigma}.$$

In an effort to reach an autonomous system it is then a good idea to define

$$(89) \quad \bar{p} = p[L(t)/N(t)]^{\sigma},$$

so that

$$(90) \quad U'(z) = \bar{p},$$

and then seek a differential equation for  $\bar{p}$  to replace (82). Take the logarithm of both sides in (89), differentiate with respect to time and

substitute from (82) and (88).

$$(91) \quad \frac{\dot{\bar{p}}}{\bar{p}} = \rho + \sigma \tau - \gamma - g'(k).$$

The system of equations (83), (91), and (90) is now again of the same form as the Ramsey model, with  $K$ ,  $C$ ,  $p$ ,  $F(K)$ , and  $\rho$  being replaced by  $k$ ,  $c$ ,  $\bar{p}$ ,  $g(k)$ , and  $(\rho + \sigma \tau) - \gamma$ , respectively. The last conditions mean that for optimality we need

$$(92) \quad \rho + \sigma \tau > \gamma,$$

with some possible cases of optimality when equality holds. It is also worth noting that, from (91), the equilibrium capital-labor ratio,  $k^\infty$ , is defined by

$$g'(k^\infty) = \rho + \sigma \tau - \gamma.$$

From (81), this can be written

$$(93) \quad f'(k^\infty) - \delta = \rho + \sigma \tau.$$

The left-hand side is thus the equilibrium net marginal productivity of capital (net of depreciation, that is) and so, in usual economic terminology, the right-hand side is an equilibrium rate of interest.

Remark 1. The existence condition (92) amounts to saying that the equilibrium rate of interest exceeds the rate of growth.

Remark 2. The equilibrium rate of interest is higher, the higher the rate of technological progress. Notice also that if  $\tau = 0$ , then the entire equilibrium does not depend in any way on the felicity function but only on the production function and the utility rate of discount,  $\rho$ . With technological progress, on the other hand, this ceases to be true; other things being equal, the marginal productivity of capital is higher

(and therefore the capital-labor ratio,  $k^{\infty}$ , is smaller) the higher  $\sigma$ , i.e., the more rapidly the individual becomes surfeited with goods.

Remark 3. Note also that  $c$  is consumption per effective worker, not consumption per capita. As the optimal path converges,  $c$  converges to a limit; but since  $L/N$  increases at the constant rate  $\tau$ , it follows that asymptotically consumption per capita will grow exponentially at the rate  $\tau$ .

## Lecture 5

### Optimal Growth in a Dual Economy

It is a common hypothesis among economists that in underdeveloped countries there exist side-by-side two economic systems, one advanced and the other backward. The economic significance of this separation is that workers in the advanced economy receive a wage which may be much higher than anything received in the backward sector. At the same time, it is assumed that these workers save nothing, so that any capital accumulation must come out of the surplus of output over wage payments. For simplicity, assume there is no relevant product at all in the backward sector. It still may not be optimal for the economy to have full employment of the labor force in the advanced section; each additional worker creates more product, on the one hand, and a claim to a fixed portion of that product on the other. Thus capital accumulation might be lower under full employment than with some unemployment.

For simplicity, it is assumed here that the population and available labor force are constant and that there is no technological progress; generalization in these directions can easily be carried out by the methods of the last lecture. The following discussion is based on the work of Marglin [1966] and Dixit [1967]. The Ramsey model is modified by adding one instrument and two constraints. The additional instrument is the amount of labor to be employed,  $L$ ; the additional constraints are that there is a fixed parameter,  $w$  (wage rate in terms of goods), such that,

$$(94) \quad C - wL \geq 0,$$

and that the amount of labor employed not exceed the fixed amount available.

$$(95) \quad \bar{L} - L \geq 0.$$

Otherwise, the Ramsey conditions remain:

$$(24) \quad \text{maximize} \int_0^{\infty} e^{-\rho t} U[C(t)] dt,$$

$$(28) \quad \dot{K} = I,$$

$$(66) \quad F(K, L) - C - I \geq 0;$$

(66) is substituted for (29) since the labor force is a variable of the problem; the function  $F$  is assumed to satisfy (65).

The Lagrangian can be written,

$$(96) \quad U(C) + pI + q_1 [F(K, L) - C - I] + q_2 (C - wL) + q_3 (\bar{L} - L).$$

Equate to zero the derivatives of the Lagrangian with respect to the three instruments,  $C$ ,  $I$ , and  $L$ .

$$U'(C) = q_1 - q_2, \quad p = q_1 F_L(K, L) = q_2 w + q_3,$$

or,

$$(97) \quad U'(C) = p - q_2,$$

$$(98) \quad p F_L = q_2 w + q_3$$

where  $F_L = \partial F / \partial L$ . Of course,

$$(99) \quad q_2 \geq 0, \quad q_2(C - wL) = 0; \quad q_3 \geq 0, \quad q_3(\bar{L} - L) = 0.$$

Since the constraint (66) is certainly effective, (28) and (66)

imply,

$$(100) \quad \dot{K} = F(K, L) - C.$$

The auxiliary equation, as before, is

$$(101) \quad \dot{p}/p = \rho - F_K(K, L).$$

From (100) and (101), at an equilibrium,

$$(102) \quad F_K(K^\infty, L^\infty) = \rho, \quad C^\infty = F(K^\infty, L^\infty).$$

To be an equilibrium of this system, however, (94) must be satisfied.

Since  $F$  is homogeneous of degree 1, it is easy to prove that  $F_K$  is homogeneous of degree 0; the first equation in (102) can therefore be

solved for  $K^{\infty}/L^{\infty}$ . Write the second equation as,

$$C^{\infty}/L^{\infty} = F(K^{\infty}/L^{\infty}, 1),$$

since  $F(K, L)$  is homogeneous of degree 1. We will assume then that,

$$(103) \quad C^{\infty}/L^{\infty} > w.$$

Then the constraint (94) is not binding at equilibrium, and  $q_2^{\infty} = 0$ . Then, from (98),  $q_3^{\infty} > 0$ , so that (95) is binding, i.e., there is full employment. Thus, for  $K$  in the neighborhood of  $K^{\infty}$ , the optimal path is identical with that for the Ramsey problem. Since, in the Ramsey problem,  $p$  is a decreasing function of  $K$ , and therefore  $C$  is an increasing function of  $K$ , it follows that the constraint (94) is fulfilled and ineffective for  $K \geq K^{\infty}$ . It follows that there is  $\bar{K} < K^{\infty}$  such that the optimal solution for the dual economy coincides with the Ramsey solution in the interval  $< \bar{K} + \infty >$ , which will be termed interval I.

$\bar{K}$  is defined by the condition that (94) becomes effective there. Since  $p(K)$  is the same as for the Ramsey solution for  $K \geq \bar{K}$ , it is now known for  $K = \bar{K}$ . Also,  $q_3(\bar{K}) = p(\bar{K}) F_L(\bar{K}, \bar{L}) > 0$ . As  $K$  decreases below  $\bar{K}$ , it must be that  $q_3$  remains positive, at least for some interval, while  $q_2$  rises above 0. Then constraints (94) and (95) are both effective in an interval to the left of  $K = \bar{K}$  — termed interval II, in which  $C = w\bar{L}$ ,  $\dot{K} = F(K, \bar{L}) - w\bar{L}$ , so that, from (101),

$$(103) \quad dp/dK = p[p - F_K(K, \bar{L})] / [F(K, \bar{L}) - w\bar{L}] \text{ in interval II.}$$

Since  $p(\bar{K})$  is known, this equation can be solved rather easily for smaller values of  $K$ .

Also in interval II,  $q_2 = p - U'(w\bar{L})$ , from (97), so that, from (98),

$$(104) \quad q_3 = p[F_L(K, \bar{L}) - w] + wU'(w\bar{L}).$$

Thus the lower end of interval II is defined by the condition  $q_3 = 0$ .

Since  $F(0,L) = 0$ , all  $L$ , by (65), there exists  $K_1$  so that,  
 $F(K_1, \bar{L}) = w\bar{L}$ .

As  $K$  approaches  $K_1 + 0$ , the denominator of (103) is asymptotically equivalent to  $F_K[K_1, \bar{L}] (K - K_1)$ , so that clearly  $p(K)$  approaches infinity. Also, from Euler's theorem on homogeneous functions,

$$w\bar{L} = F(K_1, \bar{L}) = F_L(K_1, \bar{L}) \bar{L} + F_K(K_1, \bar{L}) K_1 > F_L(K_1, \bar{L}) \bar{L},$$

so that  $F_L(K_1, \bar{L}) < w$ . The first term of (104) then approaches  $-\infty$ , while the second is constant. Hence,  $q_3(K) = 0$  for some  $K > 0$ .

Interval III is the interval  $<0, K>$ . In this interval, the full employment condition, (95), ceases to be binding, and  $q_3 = 0$ . From (94), (97), and (98), we deduce,

$$(105) \quad U'(wL) = p \{1 - [F_L(K, L)/w]\}.$$

which defines  $L$  as a function of  $K$  and  $p$ . The basic differential equation takes the form in interval III,

$$(106) \quad dp/dK = p[p - F_K(K, L)]/[F(K, L) - wL].$$

It is to be noted that  $dL/dK > 0$  in this interval (the more capital, the more labor can be employed). This means that, as we push the solution to lower values of  $K$ , the full employment constraint will never become binding again. To see that  $dL/dK > 0$  in interval III, first note, from (105) that

$$1 - [F_L(K, L)/w] > 0 \text{ in interval III.}$$

Differentiate (105) totally with respect to  $K$  and group terms.

$$(107) \quad [U''(wL) w + (p/w) F_{LL}(K, L)] (dL/dK) = - (p/w) F_{LK}(K, L) + [1 - [F_L(K, L)/w]] (dp/dK).$$

From the concavity of  $U$  and  $F$ , it follows that  $U'' < 0$ ,  $F_{LL} < 0$ .

Since  $F_L$  is homogeneous of degree 0,

$$F_{LL} L + F_{LK} K = 0,$$

by Euler's theorem; but since  $F_{LL} < 0$ , and  $L, K > 0$ ,  $F_{LK} > 0$ .

It is then easy to calculate, from (107),

(108) if  $dp/dK < 0$ , then  $dL/dK > 0$  in interval III.

Since  $\underline{K} < K^\infty$ ,  $F_K(\underline{K}, \bar{L}) > F_K(K^\infty, \bar{L}) = \rho$ , so that  $p'(\underline{K}) < 0$ . Suppose  $p'(K^*) = 0$ , for some  $K^*$ ,  $0 < K^* < \underline{K}$ . Take the largest such. Then  $p'(K) < 0$ ,  $K^* < K \leq \underline{K}$ , so that  $F_K(K, L) > \rho$  in that interval, or,

$$K/L < K^\infty/\bar{L} \text{ for } K^* < K \leq \underline{K},$$

while  $K^*/L^* = K^\infty/\bar{L}$ . Since  $F_{LK} > 0$ ,  $F_L$  increases with  $K$  for fixed  $L$ ; but since  $F_L$  is a function of  $K/L$ ,  $F_L$  increases with  $K/L$ . Hence,

$F_L(K^*, L^*) > F_L(K, L)$  for  $K$  in a right-hand neighborhood of  $K^*$ , where, it will be recalled,  $L$  is a function of  $K$  defined by (105), and  $L^*$  is its value at  $K = K^*$ . Therefore,

$$dF_L/dK \leq 0 \text{ at } K = K^*.$$

But,  $dF_L/dK = F_{LL} (dL/dK) + F_{LK}$ . Compute  $dL/dK$  from (107), and recall that  $dp/dK = 0$  at  $K = K^*$ . Then,

$dF_L/dK = F_{LK} (K^*, L^*) U''(wL^*) w / [U''(wL^*) w + (p/w) F_{LL} (K^*, L^*)] > 0$  at  $K=K^*$ , a contradiction. Hence,  $p'(K) < 0$  for  $0 < K < \underline{K}$ ; by (108)  $L$  is an increasing function of  $K$  in interval III (capital permits employment), and consumption is proportional to  $L$ .

## REFERENCES

- ARROW, K. J., L. HURWICZ, and H. UZAWA 1961 Constraint Qualifications in Nonlinear Programming. Naval Research Logistics Quarterly 8:175-191.
- ARROW, K. J., and M. KURZ 1967 Optimal Growth with Irreversible Investment in a Ramsey Model. Technical Report No. 1 (NSF GS-1440), Institute for Mathematical Studies in the Social Sciences, Stanford University, California.
- BELLMAN, R. 1957 Dynamic Programming. Princeton, N. J.: Princeton University Press.
- DIXIT, A. 1967 Optimal Development in the Labor-Surplus Economy. Unpublished manuscript.
- GORMAN, W. M. 1957 Convex Indifference Curves and Diminishing Marginal Utility. Journal of Political Economy 65:40-50.
- HALKIN, H. 1964 On the Necessary Conditions for Optimal Control of Nonlinear Systems. Journal d'analyse mathématique 12:1-82.
- KOOPMANS, T. C. 1965 On the Concept of Optimal Economic Growth. In Study Week on The Econometric Approach to Development Planning, pp. 225-87. Amsterdam: North-Holland.
- KUHN, H. W., and A. W. TUCKER 1951 Non-linear Programming. In J. Neyman (ed.) Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, pp. 481-92. Berkeley and Los Angeles: University of California Press.
- KURZ, M. 1965 Optimal Economic Growth and Wealth Effects. Technical Report 136, Institute for Mathematical Studies in the Social Sciences, Stanford University, California.

- KURZ, M. 1967 The General Instability of a Class of Competitive Growth Processes. Unpublished manuscript, Institute for Mathematical Studies in the Social Sciences, Stanford University, Stanford, California.
- MANGASARIAN, O. L. 1966 Sufficient Conditions for the Optimal Control of Nonlinear Systems. SIAM Journal on Control 4:139-52.
- MARGLIN, S. A. 1966 Industrial Development in the Labor-Surplus Economy: An Essay in the Theory of Optimal Growth. Unpublished manuscript.
- MASSE, P. 1946 Les réserves et la régulation de l'avenir. Paris: Herman, 2 vol.
- MIRRELES, J. A. 1967 Optimum Growth when the Technology is Changing. Review of Economic Studies 34:95-124.
- PONTRYAGIN, L. S., V. G. BOLTYANSKII, R. V. GAMKRELIDZE, and E. F. MISCHENKO 1962 The Mathematical Theory of Optimal Processes. New York and London: Interscience.
- RAMSEY, F. P. 1928 A Mathematical Theory of Savings. Economic Journal 38:543-59.
- SAMUELSON, P. A. 1947 The Foundations of Economic Analysis. Cambridge, Mass.: Harvard University Press.
- TINBERGEN, J. 1952 On the Theory of Economic Policy. Amsterdam: North-Holland.
- VON WEIZSÄCKER, C. C. 1965 Existence of Optimal Programs of Accumulation for an Infinite Time Horizon. Review of Economic Studies 32:85-104.

**MATHEMATICAL ECONOMICS**

by

**DAVID GALE**

at the

**American Mathematical Society Summer Seminar**

on the

**Mathematics of the Decision Sciences**

**Stanford University**

**July - August 1967**

**Analysis of a One Good Model of Economic Development**

by

**David Gale and W. R. Sutherland**

These notes have been prepared by the first author above for presentation at the 1967 Summer Seminar of the American Mathematical Society on the Mathematics of the Decision Sciences, at Stanford University. The notes present the more basic results and techniques of the theory in the first three sections. The last section applies these techniques to derive some new results obtained jointly by the two authors during the past year. Acknowledgement is here made to the National Science Foundation which is supporting the Summer Seminar as well as the preparation of these notes, and also to the Logistics and Mathematical Statistics Branch of the Office of Naval Research for support of some of the research reported on here.

## Analysis of a One Good Model of Economic Development

### 1. Introduction

In these notes we are going to analyze an idealized, or better, an imaginary economy in which there is only one good. This good can be used for two purposes; (A) it can be consumed, thus creating satisfaction or utility for the people who consume it, or (B) it can be invested, in which case it creates additional amounts of itself. We will be concerned with an operation of this economy throughout time and therefore the problem at each instant will be to decide how much to consume and how much to invest in order to maximize utility throughout time in some suitably defined sense.

What is the purpose in considering this sort of imaginary situation which bears little resemblance to any actual economy, living or dead? The answer is that in analyzing this model we shall run into certain mathematical and economic techniques which turn out to be basic not only for the study of this make-believe economy but also for the more realistic (but more complicated) models which may come up in practice. Our aim is thus to isolate this technique in a simple context. The technique we refer to is what economists describe as

the use of a price system and what mathematicians refer to as the method of dual variables. By whatever name one calls it, this subject is the central one both in economic analysis and modern optimization theory. Mathematically it enables one to answer such questions about optimal development programs as: do they exist? Are they unique? What are their qualitative properties? Economically it allows one to give a competitive market interpretation to these optimal paths along which, it turns out, producers are maximizing profits and consumers are maximizing utility subject to their budgetary limitations.

The above is a rough preview of what will be found in the rest of these notes. The overture is now ended and the show will begin.

## 2. The Model; Finite Time Horizon

The model will involve a single commodity which we will refer to as "goods". It is described by two functions, a production function  $f_t(x)$  and a utility function  $u_t(c)$  where  $f_t(x)$  is the amount of goods produced at time  $t+1$  from an investment of  $x$  units of goods at time  $t$ , and  $u_t(c)$  is the satisfaction gained by consuming  $c$  units of goods at time  $t$ . The domain of  $t$  is the

non-negative integers and that of  $x$  and  $c$  the non-negative reals.

DEFINITION 1. A program with initial stocks  $s$  is a sequence of pairs  $\langle x_t, c_t \rangle$  finite or infinite such that

$$(2.1) \quad c_0 = s - x_0$$

$$(2.2) \quad c_t = f_{t-1}(x_{t-1}) - x_t \quad \text{for } t > 0.$$

If  $\langle x_t, c_t \rangle$  is a program the corresponding utility sequence is given by  $\langle u_t(c_t) \rangle$ .

Clearly, conditions (1) and (2) state that the sum of consumption and investment in period  $t$  is equal to the amount produced in the previous period. If the sequence  $\langle x_t, c_t \rangle$  is finite with  $t = 1, \dots, T$  then it is called a T-period program and if  $f_T(x_T) = s'$  we refer to the program as a T-period program with initial stocks  $s$  and final stocks  $s'$  or, more briefly, a T-period program from  $s$  to  $s'$ . The value of such a program is  $\sum_{t=0}^T u_t(c_t)$ .

DEFINITION 2. A T-period program from  $s$  to  $s'$  is called optimal if it has maximum value among all such programs.

Although our principal interest will be in infinite rather than finite programs it will be necessary first

to develop the basic properties of finite optimal programs.

We now introduce the central concept of these notes.

DEFINITION 3. The program  $\langle x_t, c_t \rangle$  is called competitive if there exist non-negative numbers (prices)  $p_t$  such that

- (A)  $u_t(c) - p_t c$  is maximized at  $c_t$  for all  $t$ ,
- (B)  $p_{t+1} f_t(x) - p_t x$  is maximized at  $x_t$  for all  $t$ .

These conditions have an important economic interpretation. Regarding  $p_t$  as prices we see that  $p_t x$  is the cost of investing  $x$  units at time  $t$ , while  $p_{t+1} f_t(x)$  is the return or value of  $f_t(x)$  units at time  $t+1$ . The difference, therefore, represents profit and condition (B) requires that investment be chosen at each time  $t$  so as to maximize profits.

To motivate condition (A) we note from (2.2) that

$$p_t c_t = p_t (f_{t-1}(x_{t-1}) - x_t)$$

and the right hand side here might be thought of as disposable income since it represents the value of goods just produced minus the cost of goods to be invested. If we then require consumers to spend no more than the amount  $p_t c_t$  (budget constraint) condition (A) says

that consumers will then consume so as to maximize their utility subject to this constraint.

The following simple result is the starting point for the theory.

**THEOREM 1.** If  $\langle x_t, c_t \rangle$  is a T-period program from  $s$  to  $s'$  which is competitive, then it is optimal.

Proof. Let  $(p_t)$ ,  $t = 1, \dots, T+1$  be the competitive prices and let  $\langle x'_t, c'_t \rangle$  be any other program from  $s$  to  $s'$ . Then from (A) and (2.1) and (2.2)

$$u_0(c'_0) - u(c_0) \leq p_0(c'_0 - c_0) = p_0(s - x'_0) - p_0(s - x_0) = -p_0x'_0 + p_0x_0,$$

$$u_t(c'_t) - u(c_t) \leq p_t(c'_t - c_t) = p_t(f_{t-1}(x'_{t-1}) - x'_t) - p_t(f_{t-1}(x_{t-1}) - x_t) \\ \text{for } t = 1, \dots, T$$

$$\text{and} \quad 0 = p_{T+1}(s' - s) = p_{T+1}f_T(x_T) - p_{T+1}f_T(x'_T),$$

and summing on  $t$  gives

$$\sum_{t=0}^T (u_t(c'_t) - u_t(c_t)) \leq \sum_{t=0}^T \{ (p_{t+1}f_t(x'_t) - p_t x'_t) - (p_{t+1}f_t(x_t) - p_t x_t) \}$$

where we have collected terms in  $x_t$ . But since each term in the sum on the right hand side above is non-positive from (B), it follows that

$$\sum u_t(c'_t) - \sum u_t(c_t) \leq 0$$

so  $\sum u_t(c_t)$  is a maximum as asserted.

What we have here shown is that competitive programs are optimal. We need a converse to this theorem and for this purpose must make some assumptions about the functions  $f$  and  $u$ . These are

(I) The function  $f$  is non-negative, concave and increasing in  $x$  (for each  $t$ ) and  $f_t(0) = 0$ .

(II) The function  $u$  is concave and increasing in  $c$ , but possibly  $u_t(0) = -\infty$ .

The last condition of (II) is important for we would like to permit functions such as  $u(c) = \log c$ ,  $c \geq 0$ . The condition  $u(0) = -\infty$  would mean that to consume nothing (starvation) is "infinitely bad". Unless otherwise stated it will be assumed henceforth that conditions (I) and (II) are satisfied.

We now recall the fundamental mathematical result needed for this work (which may well be the fundamental result of all optimization theory), namely the Kuhn-Tucker Theorem. We can get by with the following weak form:

Kuhn-Tucker Theorem. Let  $u(x)$  and  $f_1(x)$ ,  $i = 1, \dots, m$ , be convex functions defined on a convex set  $X$  and let  $\bar{x}$  minimize  $u(x)$  in  $X$  subject to

$$(2.3) \quad f_i(x) \leq 0, \quad i = 1, \dots, m.$$

Then if (2.3) has a strict solution there exist numbers  $p_i \geq 0$  such that

(2.4)  $u(x) - \sum p_i f_i(x)$  is minimized at  $\bar{x}$ .

Suggestion for a Do-It-Yourself Proof: Let  $Y$  be the set of all  $y = (y_1, \dots, y_m)$  such that the inequalities

$$f_i(x) \leq y_i$$

have a solution. Show that  $Y$  is convex and has  $0$  as an interior point (here we use the strict solution hypothesis). Now let  $\mu(y) = \min_{f_i(x) \leq y_i} u(x)$  and show that  $\mu$  is a convex function of  $y$ . Then use the fact that a convex function  $\varphi$  has a support at every interior point  $\bar{x}$  of its domain, i.e. there is a linear function  $p \cdot x$  such that  $p \cdot (x - \bar{x}) \leq \varphi(x) - \varphi(\bar{x})$  for all  $x$  in  $X$ . The support of  $\mu$  at  $0$  is the  $p$  we are looking for.

We can now get the desired converse for Theorem 1.

**THEOREM 2.** Let  $\langle \bar{x}_t, \bar{c}_t \rangle$  be an optimal program from  $s$  to  $s'$  and assume

(2.5)  $\bar{c}_t > 0$  for at least one  $t$ .

Then  $\langle \bar{x}_t, \bar{c}_t \rangle$  is competitive.

Remark. Without (2.5) the Theorem would not be true. Suppose  $f_t(x) = px$  for some fixed  $p$  (i.e.  $f$  is linear) and suppose  $u(c) = \log c$ ,  $s = 1$ ,  $s' = p^T$ .

Then clearly the only program from  $s$  to  $s'$  is  
 $\langle x_t, c_t \rangle = \langle p^t, 0 \rangle$ , but condition (A) requires there  
 exist  $p_t$  such that

$$\log c - p_t c = \max \text{ at } c = 0$$

and clearly no such  $p_t$  exist. The fact that  $\log 0 = -\infty$   
 is not crucial here. The same situation would occur for  
 $u(c) = \sqrt{c}$ . The difficulty comes from the fact that the  
slope of  $u$  is infinite at  $c = 0$ .

Proof. Replace conditions (2.1) and (2.2) by

$$\begin{aligned} c_0 + x_0 - s &\leq 0 \\ (2.6) \quad c_t + x_t - f_{t-1}(x_{t-1}) &\leq 0 \quad t = 1, \dots, T \\ s' - f_T(x_T) &\leq 0 \end{aligned}$$

Now clearly  $\langle \bar{x}_t, \bar{c}_t \rangle$  satisfies (2.6) and it also  
 maximizes  $\sum u_t(c_t)$ , since each  $u_t$  is non-decreasing  
 in  $c$ . Hence the Kuhn-Tucker Theorem applies provided  
 we can show that (2.6) has a strict solution. Assuming  
 this for the moment we obtain numbers  $p_t \geq 0$ ,  $t = 0, \dots,$   
 $T + 1$  such that

$$(2.7) \quad \sum_{t=0}^T u_t(c_t) - p_0(c_0 + x_0) - \sum_{t=1}^T p_t[(c_t + x_t) - f_{t-1}(x_{t-1})] + p_{T+1}f_T(x_T)$$

is maximized at  $\langle \bar{x}_t, \bar{c}_t \rangle$ . Rearranging (2.7) gives

$$(2.8) \quad \sum_{t=0}^T (u_t(c_t) - p_t c_t) + \sum_{t=0}^T (p_{t+1} f_t(x_t) - p_t x_t)$$

is maximized at  $\langle \bar{c}_t, \bar{x}_t \rangle$ , but note that the terms of (2.8) are independent, hence (2.8) is maximized at  $\langle \bar{x}_t, \bar{c}_t \rangle$  if and only if  $u_t(c_t) - p_t c_t$  is maximized at  $\bar{c}_t$  and  $p_{t+1} f_t(x_t) - p_t x_t$  is maximized at  $\bar{x}_t$  for all  $t$ , and these are precisely conditions (A) and (B).

To show that (2.6) has a strict solution we consider the new program  $\langle \bar{x}_t, 0 \rangle$  and note that we have

$$(2.9) \quad \begin{aligned} \bar{x}_0 - s &\leq 0 \\ \bar{x}_t - f_t(\bar{x}_{t-1}) &\leq 0 \\ s' - f_T(\bar{x}_T) &\leq 0 \end{aligned}$$

and at least one of the above inequalities is strict by assumption (2.5). We therefore reduce the problem to the following:

**LEMMA 1.** If (2.9) has a solution with one strict inequality then it has a strict solution.

Proof. Induction on  $T$ . If  $T = 0$  we have

$$\begin{aligned} x_0 - s &\leq 0 \\ s' - f(x_0) &\leq 0 \end{aligned}$$

If  $x_0 - s < 0$  then by slightly increasing  $x_0$  if

necessary we can assure that  $s' - f(x_0) < 0$  also, since  $f$  is increasing. If  $s' - f(x_0) < 0$  then by slightly decreasing  $x_0$  we can assure  $x_0 - s < 0$  as well.

Now suppose one of the inequalities (2.9) is strict for some  $t_0 > 0$ . Then by induction hypothesis there is a solution  $x_t^i$  giving strict inequality for all but the first inequality, and to get a strict solution we slightly decrease  $x_0^i$  if necessary. In the other case we have  $x_0 - s < 0$  so inductively there is a solution  $x_t^i$  satisfying all but the last inequality strictly and this will be satisfied too by a slight increase in  $x_T^i$ .

The fact that the class of optimal and competitive programs are identical is of economic interest in itself as it shows that if the "prices are right" optimality is attained by allowing producers and consumers to act purely selfishly and maximize profits and utility respectively. We shall now show how the price theorem can be used to gain qualitative information about the nature of optimal programs. For the rest of this section we will assume that the functions  $f$  and  $u$  are independent of the time.

**DEFINITION 4.** The function  $f$  will be called productive for any  $x \geq 0$ ,  $h > 0$ ,  $f(x+h) > f(x) + h$ . In words,

increasing the input by some amount will increase the output by more than that amount. For  $f$  differentiable this is equivalent to  $f'(x) > 1$ .

**THEOREM 3.** If  $f$  is productive then for any competitive program  $\langle x_t, c_t; p_t \rangle$

- (a) prices  $p_t$  are positive and decreasing in  $t$ .
- (b) consumption  $c_t$  (and hence utility) is non-decreasing in  $t$ .
- (c) stocks  $x_t$  are non-decreasing up to some time  $t_0$  and decreasing thereafter.

Proof. (a) We first note from (A) that  $c_t$  maximizes  $u(c) - p_t c$ . This shows that  $p_t > 0$  since otherwise  $u$ , being increasing, would have no maximum. Next, from (B)

$$p_{t+1}f(x_t) - p_t x_t \geq p_{t+1}f(x) - p_t x \text{ for all } x \geq 0$$

$$\text{or } \frac{p_t}{p_{t+1}} \geq \frac{f(x) - f(x_t)}{x - x_t} \text{ for all } x > x_t$$

but since  $f$  is productive the right hand side above is greater than 1, hence  $p_{t+1} < p_t$ .

(b) From (A)

$$u(c_t) - u(c_{t+1}) \geq p_t(c_t - c_{t+1})$$

$$u(c_{t+1}) - u(c_t) \geq p_{t+1}(c_{t+1} - c_t)$$

hence

$$0 \geq (p_t - p_{t+1})(c_t - c_{t+1})$$

(this relation is sometimes called La Chatelier's principle, I think). But from (a)  $p_t - p_{t+1} > 0$  hence  $c_t - c_{t+1} \leq 0$  as asserted.

(c) It will suffice to show that if  $x_t < x_{t-1}$  then  $x_{t+1} < x_t$ . Now

$$x_{t+1} - x_t = f(x_t) - f(x_{t-1}) - (c_{t+1} - c_t) \leq f(x_t) - f(x_{t-1}) \text{ from (b).}$$

From (B)

$$p_t f(x_t) - p_{t-1} x_t \leq p_t f(x_{t-1}) - p_{t-1} x_{t-1}$$

$$\text{so } f(x_t) - f(x_{t-1}) \leq (p_{t-1}/p_t)(x_t - x_{t-1}) < 0$$

$$\text{so } x_{t+1} - x_t < 0.$$

### 3. Infinite Programs.

The finite horizon programs are not of great interest in economic development. It is true that if one were devising, say a five year plan and had decided on the final stocks  $s'$  then it would be natural to try to solve the problem of the previous section. However, the important decision would in this case already have been

made, namely the choice of  $s'$ . The main problem in economic planning is to set reasonable goals for capital accumulation and it appears that the only way to attack this is to consider infinite programs. The first thing needed is a notion of optimality.

DEFINITION 5. If  $\langle x_t, c_t \rangle$  and  $\langle x'_t, c'_t \rangle$  are infinite programs we say that  $\langle x_t, c_t \rangle$  overtakes  $\langle x'_t, c'_t \rangle$  if there exists a time  $T$  such that

$$\sum_{t=0}^{T'} u_t(c_t) > \sum_{t=0}^{T'} u_t(c'_t) \quad \text{for all } T' \geq T.$$

We say that  $\langle x_t, c_t \rangle$  catches up to  $\langle x'_t, c'_t \rangle$  (at infinity) if

$$\lim_{T \rightarrow \infty} \sum_{t=0}^T u(c_t) - u(c'_t) \geq 0.$$

A program will be called optimal (strongly optimal) if it catches up to (overtakes) every other program.

We remark that if it should happen that the series  $\sum_{t=0}^{\infty} u_t(c_t)$  converge for all programs (as may occur, for instance if future utilities are suitably discounted) then Definition 5 corresponds to choosing as the optimal program the one whose utility sum is greatest, just as in the finite case. However, Definition 5 is more general for, as we shall see, optimal programs in this broader sense may exist although all the utility series are

divergent. In the next section we shall give a specific rather general existence theorem for the case when  $u$  and  $f$  are independent of the time. In the present section we shall obtain infinite analogues to Theorems 1 and 2 relating optimal and competitive programs. Note in this connection that the definition of a competitive program requires no modification for the infinite case since conditions (A) and (B) carry over as given.

**THEOREM 4.** Any optimal program  $\langle x_t, c_t \rangle$  is competitive.

Proof. We first dispose of a trivial case in which  $c_t = 0$  for all but a finite number of times  $t$ . This means that all stocks are completely consumed by the end of  $T$  time periods for some  $T$ , so we are back in the finite case of Theorem 2 where the final stocks  $s'$  are zero.

In all other cases  $c_t > 0$  for infinitely many  $t$ . Note that if we truncate the program at  $t = T$  we have an optimal  $T$ -period program (with final stocks  $f_T(x_T)$ ), so for each  $T$  there exist prices  $p_t^T$  which satisfy (A) and (B) for  $t \leq T$ . Denoting by  $\Pi_t^T$  the set of all such prices one verifies that  $\Pi_t^T$  is a closed interval, possibly unbounded above, of non-negative numbers. Also  $\Pi_t^{T+1} \subset \Pi_t^T$  since if (A) and (B) are satisfied for  $t \leq T+1$  they are satisfied for  $t \leq T$ , so it remains

to show  $\Pi_t = \bigcap_{T=1}^{\infty} \Pi_t^T$  is non-empty, and this will follow from the nested interval theorem if we can show that for any  $t$  there exists  $T$  such that  $\Pi_t^T$  is bounded. We first note that  $\Pi_T^T$  is bounded if  $c_T > 0$ , for from (A) if  $p_T^T \in \Pi_T^T$  then

$$u_T(c_T) - p_T^T c_T \geq u_T(c_T/2) - p_T^T c_T/2$$

or 
$$p_T^T \leq 2(u_T(c_T) - u_T(c_T/2))/c_T .$$

Now for any  $t$  from (B)

$$p_{t+1}^T f_t(x_t) - p_t^T x_t \geq 0 \quad (\text{since } f_t(0) = 0)$$

so 
$$p_t^T \leq (f_t(x_t)/x_t) p_{t+1}^T .$$

Letting  $q_t = f_t(x_t)/x_t$  we have

$$p_t^T \leq q_t q_{t+1} \cdots q_{T-1} p_T^T$$

and this establishes the desired bound, and shows the existence of the competitive prices.

We would like now to establish some sort of analogue of Theorem 1 asserting that competitive programs are optimal, but since we do not have the concept of final stocks some additional condition will be required. Before continuing we consider a concrete example.

EXAMPLE 1. Let  $u(c) = -\frac{1}{c}$ ,  $f(x) = \rho x$ ,  $\varepsilon = 1$ ,  
where  $\rho$  is some positive constant.

Proposition 1. The sequence  $\langle x_t, c_t \rangle$  is a program  
if and only if  $\sum_{t=0}^{\infty} c_t / \rho^t \leq 1$ .

Proof. We have

$$c_0 = 1 - x_0$$

$$c_t = \rho x_{t-1} - x_t.$$

Multiplying the equation by  $1/\rho^t$  and summing gives

$$\sum_{t=0}^T c_t / \rho^t = 1 - x_T / \rho^T.$$

Conversely, let  $q_T = \sum_{t=0}^T c_t / \rho^t$  and let  $x_T = \rho^T (1 - q_T)$ .

Then  $x_0 = 1 - c_0$  and  $x_T - \rho x_{T-1} = \rho^T (1 - q_T) - \rho \rho^{T-1} (1 - q_{T-1}) =$   
 $= \rho^T (q_T - q_{T-1}) = c_T$ , so  $\langle x_T, c_T \rangle$  is a program.

Proposition 2. Competitive programs exist if and only  
if  $\rho > 1$ .

Proof. Let  $p_t$  be competitive prices. Then from (A)

$$u(c) - p_t c = -1/c + p_t c \text{ is maximized at } c_t,$$

thence  $u'(c_t) = p_t = 1/c_t^2$  or

$$(3.1) \quad c_t = 1/\sqrt{p_t}$$

and hence  $c_t$  and  $x_t$  are positive for all  $t$ . From (B)

$(p_{t+1}\rho - p_t)x$  is maximized at  $x_t$ , so  $p_{t+1} = p_t/\rho$ , hence

$$(3.2) \quad p_t = p_0/\rho^t.$$

Letting  $\sigma = \sqrt{\rho}$  we have from (3.1)  $c_t = \sigma^t/\sqrt{p_0}$  and  $c_t/\rho^t = 1/(\sqrt{p_0} \sigma^t)$ . The series  $\sum_{t=0}^{\infty} c_t/\rho^t$  will converge if and only if  $\rho > 0$ , in which case

$$(3.3) \quad \sum_{t=0}^{\infty} c_t/\rho^t = \sigma/\sqrt{p_0} (\sigma-1),$$

so  $\langle x_t, c_t \rangle$  is competitive if and only if  $c_t = 1/\sqrt{p_0} \sigma^t$  where

$$p_0 \geq (\sigma/\sigma-1)^2.$$

Proposition 3. The optimal program is  $\langle \bar{x}_t, \bar{c}_t \rangle$  where  $\bar{c}_t = \sigma^t - \sigma^{t-1}$ .

If there is any optimal program it must be competitive by Theorem 4 and clearly the best of the competitive programs (3.1) is the one for which  $p_0 = (\sigma/\sigma-1)^2$ . However, we can prove directly that this program is optimal. Let  $\langle c_t, x_t \rangle$  be any other program. From (A) and (3.2) we have

$$u(c_t) - (p_0/\rho^t)c_t \leq u(\bar{c}_t) - (p_0/\rho^t)\bar{c}_t \quad \text{or}$$

$$(3.4) \quad u(c_t) - u(\bar{c}_t) \leq p_0(c_t/\rho^t - \bar{c}_t/\rho^t) .$$

Further if for some  $t$   $\bar{c}_t \neq c_t$  then (3.4) is strict.

Summing on  $t$  gives

$$(3.5) \quad \sum_{t=1}^T u(c_t) - u(\bar{c}_t) < p_0 \sum_{t=1}^T (c_t/\rho^t - \bar{c}_t/\rho^t),$$

but from Proposition 1 and the fact that  $\sum_{t=1}^{\infty} \bar{c}_t/\rho^t = 1$ ,

it follows that the right hand side of (3.5) converges to some non-positive number and hence the left hand side must eventually become and remain negative, proving the asserted optimality.

We now prove a converse of Theorem 4.

THEOREM 5. If  $\langle \bar{x}_t, \bar{c}_t; p_t \rangle$  is competitive and

$$(3.6) \quad \lim_{t \rightarrow \infty} p_t \bar{x}_t = 0$$

then  $\langle \bar{x}_t, \bar{c}_t \rangle$  is optimal.

Proof. Let  $\langle x_t, c_t \rangle$  be any program from  $s$  and let  $\pi_t$  denote the profit from this program at prices  $\bar{p}_t$  in period  $t$ ; that is

$$\pi_t = p_t f(x_{t-1}) - p_{t-1} x_{t-1}, \quad \bar{\pi}_t = p_t f(\bar{x}_{t-1}) - p_{t-1} \bar{x}_{t-1}.$$

From (A) we have

$$\begin{aligned} (3.7) \quad & u(c_0) - u(\bar{c}_0) \leq p_0(c_0 - \bar{c}_0) = p_0(s - x_0) - p_0(s - \bar{x}_0) = -p_0(x_0 - \bar{x}_0) \\ & u(c_t) - u(\bar{c}_t) \leq p_t(c_t - \bar{c}_t) = p_t(f(x_{t-1}) - x_t) - p_t(f(\bar{x}_{t-1}) - \bar{x}_t), \quad t \geq 1, \end{aligned}$$

so

$$(3.8) \quad \sum_{t=0}^T u(c_t) - \sum_{t=0}^T u(\bar{c}_t) \leq \sum_{t=1}^T (\pi_t - \bar{\pi}_t) + p_T \bar{x}_T - p_T x_T.$$

From (B)  $\pi_t \leq \bar{\pi}_t$  so the sum on the right is non-positive and since  $p_T \bar{x}_T \rightarrow 0$  it follows that the entire right hand side becomes less than any preassigned positive number for  $T$  sufficiently large, which is the definition of optimality.

Corollary. If  $u$  is strictly concave then  $\langle \bar{x}_t, \bar{c}_t \rangle$  is strongly optimal.

Proof. In this case if  $c_t \neq \bar{c}_t$  for some  $t$  then the corresponding inequality of (3.7) becomes strict and the argument above shows that (3.8) becomes negative.

We now give an important equivalent interpretation to condition (3.6).

We define  $p_0 s$  to be the initial wealth of the economy. We define

$$W_T = p_0 s + \sum_{t=1}^T \pi_t, \quad \text{the } \underline{\text{accumulated wealth}} \text{ up to period } T$$

$$E_T = \sum_{t=0}^T p_t c_t, \quad \text{the } \underline{\text{expenditure on consumption}} \text{ up to period } T$$

$$p_T x_T = \underline{\text{value of stocks}} \text{ in period } T.$$

Then we have the following obvious identity

$$p_T x_T = W_T - E_T$$

which is obtained by multiplying the  $t^{\text{th}}$  equation of (2.1), (2.2) by  $p_t$  and adding.

In particular we have

$$(3.9) \quad E_T \leq W_T$$

which is an obvious budget inequality, stating that expenditure on consumption cannot exceed accumulated wealth. Condition (3.6) now becomes

$$(3.10) \quad \lim_{T \rightarrow \infty} (W_T - E_T) \equiv 0$$

so that "at infinity" all wealth has been used up in consumption.

The condition seems like a reasonable one for optimality. However, it is not a necessary condition. One can show that for cases in which the function  $f$  is not productive, so that eventually  $f(x) < x$ , then  $p_T x_T$  converges to some positive value rather than to zero.

We call a program efficient if it satisfies (3.10).

#### 4. The Time Independent Case

In this section we confine ourselves to the case where  $f$  and  $u$  are independent of time. We need one more assumption which is a strengthening of Definition 4.

DEFINITION 5. The function  $f$  is strongly productive if there is a constant  $\rho > 1$  such that

$$f(x + h) > f(x) + \rho h.$$

For  $f$  differentiable this is equivalent to  $f'(x) > \rho$ .

EXISTENCE THEOREM. If  $f$  is strongly productive there exists an optimal program if and only if  $u$  is bounded.

This theorem was originally proved by D. McFadden for the case of  $f$  a linear function. We first prove the necessity of the boundedness condition.

LEMMA 2. If  $\langle x_t, c_t; p_t \rangle$  is competitive then  $p_t \leq p_0 / \rho^t$ .

Proof. From (B)

$$p_{t+1}f(x_t) - p_t x_t \geq p_{t+1}f(x) - p_t x \quad \text{for all } x \geq 0$$

$$\text{or} \quad p_{t+1}(f(x_t) - f(x)) \geq p_t(x_t - x)$$

$$\text{or} \quad p_{t+1}/p_t \leq \frac{(x_t - x)}{f(x_t) - f(x)} \leq \frac{1}{\rho} \quad \text{for } x > x_t$$

from which the result follows.

LEMMA 3. If  $\langle x_t, c_t; p_t \rangle$  is competitive then  $p_t$  approaches 0 and  $x_t$  and  $c_t$  approach  $\infty$  monotonically.

Proof. The first assertion follows from the previous Lemma. Suppose  $(c_t)$  were bounded. Then there would exist  $\bar{c}$  such that  $c_t < \bar{c}$  and  $u(\bar{c}) - u(c_t) \geq \delta > 0$  for all  $t$ . But from (A)

$$p_t(\bar{c} - c_t) \geq u(\bar{c}) - u(c_t) > \delta \quad \text{for all } t$$

and we have seen that the left hand side above approaches zero, giving a contradiction. Since  $c_t$  becomes infinite so does  $x_t$  and monotonicity follows from Theorem 3.

THEOREM 6. If there is an optimal program then  $u$  must be bounded.

Proof. Let  $\langle x_t, c_t \rangle$  be an optimal, hence competitive, program. From (A)

$$\begin{aligned} u(c_1) - u(c_0) &\leq p_0(c_1 - c_0) = p_0(f(x_0) - x_1) - p_0(s - x_0) \\ u(c_{t+1}) - u(c_t) &\leq p_t(c_{t+1} - c_t) = p_t(f(x_t) - x_{t+1}) - p_t(f(x_{t-1}) - x_t). \end{aligned}$$

Summing from  $t = 1$  to  $T$

$$\begin{aligned} u(c_{T+1}) - u(c_0) &\leq p_T(x_T - x_{T+1}) + \sum_{t=1}^T [(p_t f(x_t) - p_{t-1} x_t) - \\ &\quad - (p_t f(x_{t-1}) - p_{t-1} x_{t-1})] + p_0(f(x_0) - s) \end{aligned}$$

but, the first term above is non-positive by Lemma 3 and the terms in the summation are non-positive from (B). Hence

$$u(c_{T+1}) \leq u(c_0) + p_0(f(x_0) - s),$$

so  $u(c_t)$  is bounded for all  $t$ , but since  $c_t \rightarrow \infty$  this means that  $u$  is bounded.

If  $u$  is bounded we establish the existence of an optimal program by taking the limit as  $T \rightarrow \infty$  of  $T$ -period programs, as follows:

Let  $P^T = \langle x_t^T, c_t^T; p_t^T \rangle$  be a  $T$ -period program which maximizes  $\sum_{t=0}^T u(c_t)$  (the final stocks in this case are zero). Now for a fixed  $t$ , the sets  $\{x_t^T\}$  and  $\{c_t^T\}$  are bounded for all  $T$ . If in addition we knew that  $\{p_t^T\}$  was bounded, then a standard "diagonal process" argument would establish the existence of a competitive program  $\bar{P}$ , a point-wise limit of the programs  $P^T$ . Our procedure will be first to prove the boundedness of  $\{p_t^T\}$  and then to show that  $\bar{P}$  is efficient and hence optimal, by Theorem 5.

We first need a fundamental inequality.

LEMMA 4. If  $\langle x_t, c_t; p_t \rangle$  is competitive then

$$\sum_{t=T_1}^T p_t c_t \leq p_{T_1} c_{T_1} + \rho[u(c_T) - u(c_{T_1})]/(\rho-1)$$

Proof. From (A)

$$\begin{aligned}
 u(c_T) - u(c_{T_1}) &= \sum_{T_1}^{T-1} (u(c_{t+1}) - u(c_t)) \leq \sum_{T_1}^{T-1} p_t (c_{t+1} - c_t) \\
 &= \sum_{T_1}^{T-1} (p_t - p_{t+1}) c_{t+1} + p_T c_T - p_{T_1} c_{T_1} \\
 &= \sum_{T_1}^{T-1} (1 - p_{t+1}/p_t) p_t c_t + p_T c_T - p_{T_1} c_{T_1}
 \end{aligned}$$

so from Lemma 2

$$u(c_T) - u(c_{T_1}) \leq (1 - \frac{1}{\rho}) \sum_{T_1} p_t c_t - p_{T_1} c_{T_1}$$

and we obtain (\*) by rearranging.

We need a simple property of bounded functions.

**DEFINITION 6.** The member  $p_c$  is a support of the function  $u$  at the point  $c$  if  $u(c) - u(\bar{c}) \leq p_c (c - \bar{c})$  for all  $c$ . (If  $u$  is differentiable then  $p_c = u'(c)$ . Note that  $p_t$  is a support of  $u$  at  $c_t$  in any competitive program.)

**LEMMA 5.** If  $u$  is bounded and  $p_c$  is a support of  $u$  at  $c$  then  $\lim_{c \rightarrow \infty} p_c c = 0$ .

Proof. Let  $\mu = \sup_{c \geq 0} u(c)$  and choose  $\bar{c}$  so that

$$u(\bar{c}) \geq \mu - \epsilon/2, \text{ hence } u(c) - u(\bar{c}) \leq \epsilon/2 \text{ for all } c \geq 0.$$

Then

$$\epsilon/2 \geq u(c) - u(\bar{c}) \geq p_c(c - \bar{c}) = p_c c(1 - \bar{c}/c)$$

so if  $c > 2\bar{c}$  then  $p_c c \leq \epsilon$ .

COROLLARY. The set of numbers  $p_c c$  is bounded.

We now get a first economic application.

THEOREM 7. If  $u$  is bounded there exists a number  $M$  such that for any competitive program  $\langle x_t, c_t; p_t \rangle$  the quantity  $E_T = \sum_{t=0}^T p_t c_t \leq M$  for all  $T$ .

Proof. Apply (\*) with  $T_1 = 0$  to get

$$E_T \leq p_0 c_0 + \rho/\rho-1[u(c_T) - u(c_0)]$$

but the right hand side is bounded by hypothesis and the preceding Corollary.

COROLLARY. If  $\langle x_t, c_t; p_t \rangle$  is an infinite competitive program then  $\sum_{t=0}^{\infty} p_t c_t$  converges.

LEMMA 6. For the programs  $P^T$  the prices  $p_0^T$  satisfy

$$p_0^T \leq M/s.$$

Proof. Since the final stocks  $x_T = 0$  in  $P^T$  inequality (3.9) becomes

$$p_0^T s + \sum_{t=1}^T \pi_t^T = E_T^T \leq M \text{ by Theorem 7,}$$

and since  $\pi_t^T \geq 0$  the result follows.

COROLLARY. The prices  $p_t^T$  satisfy

$$p_t^T \leq M/sp^T$$

Proof. Lemma 2.

THEOREM 8. There exists an infinite competitive program.

Proof. Take the point-wise limit of the programs  $p^T$  and call this limit  $\bar{p}$ . It is a standard exercise to verify that  $\bar{p}$  is a competitive program.

To complete the existence theorem we must prove that  $\bar{p}$  is efficient. Let  $\bar{E} = \sum_{t=0}^{\infty} \bar{p}_t \bar{c}_t$  which exists

by the Corollary to Theorem 7. Let  $\bar{W} = \bar{p}_0 s + \sum_{t=1}^{\infty} \bar{\pi}_t \epsilon$ .

We must show that this expression converges and that  $\bar{E} = \bar{W}$ .

It will be convenient to consider the program  $p^T$  to be infinite with the convention that for  $t > T$   $x_t = c_t = 0$ .

LEMMA 7. For any  $\epsilon > 0$  there exists  $t_\epsilon$  such that

$$\sum_{t=t_\epsilon}^{\infty} p_t^T c_t^T < \epsilon \quad \text{and} \quad \sum_{t=t_\epsilon}^{\infty} \pi_t^T < \epsilon \quad \text{for all } T.$$

Proof. From Lemma 6, we can choose  $t_1$  so that  $p_{t_1}^T$  is arbitrarily small, but as  $p_t \rightarrow 0$ , we have

$c_t \rightarrow \infty$  (Lemma 3) hence  $u(c_t) \rightarrow \mu = \sup_{c \geq 0} u(c)$  and

$p_t c_t \rightarrow 0$  since  $p_t$  is a support of  $u$  at  $c_t$  (Lemma 5)

then we can choose  $t_\epsilon$  so that  $p_{t_\epsilon}^T c_{t_\epsilon}^T \leq \epsilon/2$  and

$\rho/\rho-1(\mu - u(c_{t_\epsilon}^T)) \leq \epsilon/2$  for all  $T$ . Now apply (\*) and

we have

$$\sum_{t=t_\epsilon}^T p_t^T c_t^T \leq p_{t_\epsilon}^T c_{t_\epsilon}^T + [u(c_T) - u(c_{t_\epsilon})] \leq \epsilon \text{ for all } T.$$

Finally, by (3.9),  $\sum_{t=0}^{t_\epsilon-1} p_t^T c_t^T \leq p_0^T s + \sum_{t=1}^{t_\epsilon-1} \pi_t^T$  so

$$\sum_{t=t_\epsilon}^{\infty} \pi_t^T = \sum_{t=t_\epsilon}^T \pi_t^T \leq \sum_{t=t_\epsilon}^T p_t^T c_t^T = \sum_{t=t_\epsilon}^{\infty} p_t^T c_t^T \leq \epsilon.$$

**THEOREM 9.** The program  $\bar{P}$  is efficient.

Proof. Let  $E^T = \sum_{t=0}^{\infty} p_t^T c_t^T$  and let  $W^T = p_0^T s + \sum_{t=1}^{\infty} \pi_t^T$

and let  $\bar{E} = \sum_{t=0}^{\infty} \bar{p}_t \bar{c}_t$ . Now for  $T', T > t_\epsilon$  it follows

from Lemma 7 that  $E^{T'} - E^T \leq \epsilon$  and  $W^{T'} - W^T < \epsilon$ ,

so  $(E^T)$  and  $(W^T)$  are Cauchy sequences and converge

to their point-wise limits  $\bar{E}$  and  $\bar{W}$ . But  $E^T = W^T$

for all  $T$ , hence  $\bar{E} = \bar{W}$ , completing the proof.

APPENDIX. The Case of More than One Good.

In these notes the entire analysis of optimal programs has been based on the use of competitive prices, and the existence of these prices therefore played a key role. To establish their existence we were at some pains in the proofs of Theorems 4 and 8 to obtain an a priori bound on the values of prices for finite horizon programs. This boundedness requirement is no mere mathematical technicality but is quite essential to the understanding of the models. We will here illustrate this further by considering a very simple two good model in which there is an obvious optimal program which, however, is not competitive.

The model involves both a production good  $P$  and a consumption good  $Q$ , and there is a single joint process for producing both. Namely from  $x$  units  $P$  invested in period  $t$  one obtains  $\rho x$  units of  $P$  and  $x$  units of  $Q$  in period  $t + 1$ . Assuming initial stock of  $P$  is 1 a program  $\langle x_t, c_t \rangle$  must satisfy

$$x_0 \leq 1, \quad x_t \leq \rho x_{t-1} \quad \text{and} \quad c_t \leq x_t \quad \text{for all } t.$$

The inequalities here simply have the meaning that one can throw away either production or consumption goods.

Now, it is perfectly clear that by any reasonable definition of optimality the only optimal program is

$x_t = c_t = \rho^t$  since any other program involves needless throwing away. It also follows from the Kuhn-Tucker Theorem that every T-period optimal program is competitive for any utility function  $u$ . However

**THEOREM 10.** If  $\rho > 1$  and the utility function  $u$  is unbounded then the optimal program is not competitive.

Proof. We must first write down the competitive conditions. Let  $p_t$  and  $q_t$  be the prices of  $P$  and  $Q$  in period  $t$ . Condition (A) then remains

$$(A') \quad u(c) - q_t c \text{ is maximized at } c_t,$$

and the profit condition (B) at time  $t$  is clearly

$$(B') \quad q_t x + p_{t+1} \rho x - p_t x \text{ is maximized at } x_t.$$

Suppose now that  $\langle \rho^t, \rho^t \rangle$  is optimal. Then from (B') we must have

$$(1) \quad q_t = (p_t - \rho p_{t+1})$$

and from (A')

$$u(\rho^t) - u(\rho^{t+1}) \geq q_t(\rho^t - \rho^{t+1}) = (1-\rho)\rho^t(p_t - \rho p_{t+1}).$$

Summing from  $t = 0$  to  $T-1$  gives

$$u(1) - u(\rho^T) \geq (1-\rho)(p_0 - \rho^T p_T) \text{ or}$$

$$(\rho-1)p_0 \geq u(\rho^T) - u(1) + (\rho-1)\rho^T p_T \geq u(\rho^T) - u(1) \text{ for all } T$$

but if  $u$  is unbounded this is impossible since  $p_0$  would have to be infinite.

**Lectures on**  
**COMPUTATIONAL ASPECTS OF CONTROL THEORY**

**by**  
**J. B. ROSEN**

**at the**  
**American Mathematical Society Summer Seminar**  
**on the**  
**Mathematics of the Decision Sciences**  
**Stanford University**  
**July - August 1967**

14

## Optimal Control and Convex Programming<sup>1</sup>

J. B. ROSEN

*University of Wisconsin*

### INTRODUCTION

The problems arising in optimal control theory are similar mathematically to those met in the calculus of variations, with additional requirements in the form of inequality constraints which must be satisfied. The subject received its initial impetus from problems arising in the area of guidance and control, and the basic results of Pontryagin *et al.* (1962) are developed from this point of view, as is much of the subsequent work on this subject (Leitmann, 1962). However, as emphasized by Bellman, Glicksberg, and Gross (1958), a continuous spectrum of problems encountered by systems analysts, operations researchers, economists, and management consultants in various phases of industrial, scientific, and military activity can be included in an appropriate formulation of control theory. Two such potentially important applications are dynamic economic models (Usawa, 1964) and long-range capital investment studies.

As a greatly simplified example of the latter application, suppose that the control  $u(t)$  is the rate of investment at time  $t$ . The state of the system  $x(t)$  is described by the quantity of the  $i$ th product  $x_i(t)$  produced by time  $t$ . The  $x_i(t)$  are determined for any given  $u(t)$  by the system of differential equations

$$\dot{x} = f(x, u, t), \quad x(0) = x_0, \quad t \in [0, T].$$

The rate of investment is, of course, nonnegative and also may not exceed a specified upper bound, so that  $0 \leq u(t) \leq \alpha$ . Furthermore, it is required that the production schedule satisfy the state constraints  $p_i(t) \leq x_i(t) \leq q_i(t)$ ,

<sup>1</sup> Sponsored in part by NASA grant NsG 565 and in part by the Mathematics Research Center, United States Army, Madison, Wisconsin under Contract No. DA-11-022-ORD-2069.

where the  $p_i(t)$  and  $q_i(t)$  are specified, and that this be done so as to minimize the total discounted investment

$$\rho[u] = \int_0^T e^{-\rho t} u(t) dt$$

over a finite time  $T$ . Because of the presence of the state constraints, this problem is of a type which is difficult both theoretically and computationally (see Berkovits, 1962, and Pontryagin *et al.*, 1962, chap. 6). In actual practice the investment decisions would not be made continuously but rather at discrete intervals, say, once a month. This is typical of a dynamic process which can be formulated as continuous but which is more usefully considered as discrete, since this gives both a more realistic model and a computational method of solution.

The two important questions to be answered are:

1. Will any admissible ( $0 \leq u(t) \leq \alpha$ ) investment program satisfy the production constraints? That is, does an admissible control exist?
2. If there are admissible controls, how do we find one which is optimal?

The remainder of this paper is devoted to answering these two questions for a general class of discrete optimal control problems.

Some of the material here is based on parts of an earlier report (Rosen, 1964). The author has also had the benefit of several discussions with J. Abadie, whose work in this area has been most stimulating.

#### DISCRETE PROBLEM WITH STATE CONSTRAINTS

It will be useful to give a further motivation for the approach taken here for the solution of optimal control problems. Such problems fall naturally into two classes depending on their initial formulation, namely, continuous and discrete. In general, we will solve the continuous problems on a digital computer; this will require the numerical integration of systems of differential equations—in fact, a discrete approximation to the continuous process. We may therefore assume, at least for computational purposes, that we will always be dealing with discrete problems.

To be specific, we will consider a discrete problem as follows: Let  $x_i \in E^n$  represent the state vector at time  $t_i$  ( $i = 0, 1, \dots, m$ ) and  $u_i \in E^r$  the corresponding control vector for  $i = 0, 1, \dots, m-1$ . The initial value  $x_0$  is specified, and we wish to determine the vectors  $x_i$  and  $u_i$  so as to minimize

$$\sum_{i=0}^{m-1} c(x_i, u_i), \quad (1)$$

where the  $x_i$  and  $u_i$  must satisfy the recursion relation

$$x_{i+1} - x_i = f(x_i, u_i), \quad i = 0, 1, \dots, m-1, \quad (2)$$

where each  $u_i$  must be selected from a convex, compact subset  $U \subset E^r$  and where  $x_m$  must lie in a convex, compact subset  $X_m \subset E^n$ . We assume that  $\sigma(x, u)$  is a function from  $E^n \times U$  to  $E^1$  with  $\sigma \in C^1$  on  $E^n \times U$  and that  $f(x, u)$  is a function from  $E^n \times U$  to  $E^n$  with  $f \in C^1$  on  $E^n \times U$ . We assume, further, that the sets  $U$  and  $X_m$  are each specified by a system of inequality constraints, that is,

$$X_m = \{x \mid g(x) \leq 0\} \quad (3)$$

and

$$U = \{u \mid h(u) \leq 0\}, \quad (4)$$

where  $g(x)$  is a function from  $E^n$  to  $E^1$  with  $g \in C^1$  and convex on  $E^n$  and where  $h(u)$  is a function from  $E^r$  to  $E^1$  with  $h \in C^1$  and convex on  $E^r$ . The sets  $X_m$  and  $U$  are assumed to be nonempty; by the convexity of  $g(x)$  and  $h(u)$ , they are convex.

We may think of this discrete problem as arising from a finite difference approximation to the continuous problem

$$\min \int_0^T \sigma(x(t), u(t)) dt, \quad (5)$$

where

$$\begin{aligned} \dot{x} &= f(x, u), & t \in [0, T], \\ x(0) &= x_0, & x(T) \in X_m, \\ u(t) &\in U, & t \in [0, T]. \end{aligned} \quad (6)$$

The sum (1) is the simplest approximation to the integral (5) with  $\Delta t = T/m$ ,  $t_i = i\Delta t$ , and  $\sigma = \Delta t \bar{\sigma}$ . The recursion relation (2) is the simplest finite difference approximation to the differential equation (6) with  $f = \Delta t \dot{f}$ .

We may now consider the discrete problem as the minimization of a convex function on a finite-dimensional Euclidean space subject to the equality constraints (2) and the inequality constraints (3) and (4). For problems of this type, the appropriate theory is that developed by Kuhn and Tucker (1951); see also Karlin (1959) and Panagiotou (1963). For our purposes, the most convenient statement of this theory is essentially that given by Berge.

We let  $s = m(n+r)$  and denote by  $s \in E^s$  the vector  $s' = (x_1', \dots, x_m', u_0', \dots, u_{m-1}')$ , where unprimed vectors are column vectors and where the prime denotes transpose. We will call  $s$  an *admissible point* if the  $x_i$  ( $i = 1, \dots, m$ ) satisfy (2),  $x_m \in X_m$ , and  $u_i \in U$  ( $i = 0, 1, \dots, m-1$ ). Suppose we have an admissible point  $s^*$ , determined by  $x_i^*$ ,  $i = 1, \dots, m$ ,

$u_i^*, i = 0, \dots, m-1$ . We will denote by  $g_*(x_m^*)$  the  $k \times n$  Jacobian matrix of  $g(x)$  evaluated at  $x_m^*$  and by  $h_*(u_i^*)$  the  $l \times r$  Jacobian matrix of  $h(u)$  evaluated at  $u_i^*$ . We will also denote by  $\hat{g}_*(x_m^*)$  the matrix in which we have replaced by zeros the  $j$ th row of  $g_*(x_m^*)$  if the  $j$ th element of  $g(x_m^*) < 0$ . Thus, we have  $g'(x_m^*)\hat{g}_*(x_m^*) = 0$ . The matrix  $\hat{h}_*(u_i^*)$  is defined similarly for  $i = 0, 1, \dots, m-1$ , so that  $h'(u_i^*)\hat{h}_*(u_i^*) = 0$  ( $i = 0, 1, \dots, m-1$ ). We also let  $f_*(x, u)$  and  $f_*(x, u)$  denote the  $n \times n$  and  $n \times r$  Jacobian matrices of  $f$ .

An admissible direction  $\bar{z}$  at  $z^*$  is given by vectors  $\bar{x}_i$  ( $i = 1, \dots, m$ ) and  $\bar{u}_i$  ( $i = 0, \dots, m-1$ ) such that

$$\bar{x}_{i+1} - \bar{x}_i = f_*(x_i^*, u_i^*)\bar{x}_i + f_*(x_i^*, u_i^*)\bar{u}_i, \quad i = 0, \dots, m-1,$$

$$\bar{x}_0 = 0$$

and

$$\hat{g}_*(x_m^*)\bar{x}_m \leq 0,$$

$$\hat{h}_*(u_i^*)\bar{u}_i \leq 0, \quad i = 0, 1, \dots, m-1.$$

It follows that if  $y \in E^*$  is not an admissible direction at  $z^*$ , then it points outward from the set of admissible points at  $z^*$ ; that is,  $z^* + \alpha y$  is not an admissible point for every sufficiently small  $\alpha > 0$ .

The sum (1) to be minimized is given in terms of  $z$  by letting

$$\phi(z) = \sum_{i=0}^{m-1} \sigma(x_i, u_i). \quad (7)$$

We will say that an admissible point  $z^*$  is a *relative minimum* if

$$\phi(z^*) \leq \phi(z^* + \alpha \bar{z}) \quad (8)$$

for every admissible direction  $\bar{z}$  at  $z^*$  and sufficiently small  $\alpha > 0$ .

We can now state the necessary Kuhn-Tucker conditions for a relative minimum.

**Theorem 1A:** If an admissible point  $z^*$  is a relative minimum, then there exist vectors  $\lambda_i \in E^n$  ( $i = 1, \dots, m$ ), a vector  $\nu_m \geq 0$ ,  $\nu_m \in E^k$ , and vectors  $\eta_i \geq 0$ ,  $\eta_i \in E^l$  ( $i = 0, 1, \dots, m-1$ ) such that

$$\nu_m' g(x_m^*) = 0, \quad (9)$$

$$\eta_i' h(u_i^*) = 0, \quad i = 0, 1, \dots, m-1, \quad (10)$$

and such that the Lagrangian function

$$\Phi(z) = \phi(z) + \sum_{i=0}^{m-1} \lambda_{i+1}' [x_{i+1} - x_i - f(x_i, u_i)] + \nu_m' g(x_m) + \sum_{i=0}^{m-1} \eta_i' h(u_i) \quad (11)$$

has a stationary point at  $z = z^*$ , that is,

$$\Phi_z(z^*) = 0. \quad (12)$$

*Proof:* The proof is essentially that given in Kuhn and Tucker (1951) or Berge (1963) and is based on the Farkas lemma.

*Corollary:* At a relative minimum  $z^*$ , the value of the Lagrangian function  $\Phi(z)$  and the sum (1) are equal. Furthermore, the vectors  $\lambda_i$ ,  $\nu_m$ , and  $\eta_i$  must satisfy the following system of equations:

$$\lambda_{i+1} - \lambda_i = -f'_x(x_i^*, u_i^*)\lambda_{i+1} + \sigma'_x(x_i^*, u_i^*), \quad i = 1, \dots, m-1. \quad (13)$$

$$\lambda_m = -g'_x(x_m^*)\nu_m, \quad (14)$$

and

$$h'_u(u_i^*)\eta_i = f'_u(x_i^*, u_i^*)\lambda_{i+1} - \sigma'_u(x_i^*, u_i^*), \quad i = 0, 1, \dots, m-1. \quad (15)$$

*Proof:* Because of the complementary requirements (9) and (10) and the fact that the admissible point  $z^*$  satisfies (2), we have

$$\Phi(z^*) = \phi(z^*) = \sum_{i=0}^{m-1} \sigma(x_i^*, u_i^*).$$

The system (13) through (15) is equivalent to (12) and is obtained by setting to zero the partial derivative of  $\Phi$  with respect to each component of  $z$ .

It is clear from the form of (13) that this recursion relation for the  $\lambda_i$  is closely related to the usual adjoint equation for the continuous problem. The terminal value  $\lambda_m$  for the adjoint vector is specified by (14).

In Theorem 1A, necessary conditions for a relative minimum were given with no conditions on  $\sigma(x, u)$  and  $f(x, u)$  other than differentiability. We now show that if  $\sigma(x, u)$  is convex on  $E^n \times U$  and if  $f(x, u)$  is linear on  $E^n \times U$ , then the conditions are also sufficient for a global minimum.

**Theorem 2A:** Let  $\sigma(x, u)$  be convex and  $f(x, u)$  be linear on  $E^n \times U$ . If  $z^*$  is an admissible point and there exist vectors  $\lambda_i$  and nonnegative vectors  $\nu_m$  and  $\eta_i$  such that (9), (10), (13), (14), and (15) are satisfied, then  $z^*$  is a global minimum.

*Proof:* We will denote by  $Z \subset E^n$  the direct product of the sets  $x_i \in R^n$  ( $i = 1, \dots, m$ ) and  $u_i \in U$  ( $i = 0, 1, \dots, m-1$ ). Then the function  $\phi(z)$  is convex on  $Z$ , since each term is convex on  $E^n \times U$ . The first summation in (11) is linear in  $z$  and therefore also convex on  $Z$ . The remaining two terms are convex by assumption and by the fact that  $\nu_m$  and the  $\eta_i$

are nonnegative. Therefore,  $\Phi(z)$  is convex on  $Z$ . Now a stationary point of a convex function is a global minimum, so that

$$\Phi(z^*) = \min_{z \in Z} \Phi(z). \quad (16)$$

As above, we have  $\phi(z^*) = \Phi(z^*)$ . Furthermore, for every admissible point  $z$ , we have from (2) through (4) that

$$\Phi(z) \leq \phi(z). \quad (17)$$

Then, from (16) and (17),  $\phi(z^*) \leq \phi(z)$  for every admissible point  $z$ , so that  $z^*$  is a global minimum.

By means of a straightforward modification, the previous results can be extended to include the case of constraints on the state vectors  $x_i$  ( $i = 1, \dots, m-1$ ) in addition to the constraint  $x_m \in X_m$ . To show this, let us require that

$$x_i \in X_i, \quad i = 1, \dots, m, \quad (18)$$

where each  $X_i$  is a convex subset of  $E^n$ , specified in terms of convex functions  $g^i(x)$  from  $E^n$  to  $E^{k_i}$ , for  $i = 1, \dots, m$ , with  $g^m(x) = g(x)$ . We therefore have

$$X_i = \{x \mid g^i(x) \leq 0\}, \quad i = 1, \dots, m. \quad (19)$$

Note that  $X_m$  is identical to that given by (3). An admissible point is now one which satisfies (2), (18), and  $u_i \in U$ .

The extension of Theorem 1A to this state-bounded problem is given by:

**Theorem 1B:** If an admissible point  $z^*$  is a relative minimum, then there exist vectors  $\lambda_i \in E^n$  ( $i = 1, \dots, m$ ), vectors  $\nu_i \geq 0$ ,  $\nu_i \in E^{k_i}$  ( $i = 1, \dots, m$ ), and vectors  $\eta_i \geq 0$ ,  $\eta_i \in E^l$  ( $i = 0, 1, \dots, m-1$ ) such that

$$\nu_i' g^i(x^*) = 0, \quad i = 1, \dots, m, \quad (20)$$

$$\eta_i' h(u_i^*) = 0, \quad i = 0, \dots, m-1, \quad (21)$$

and such that the Lagrangian function

$$\begin{aligned} \Phi(z) = \phi(z) + \sum_{i=0}^{m-1} \lambda_{i+1}' [x_{i+1} - x_i - f(x_i, u_i)] \\ + \sum_{i=1}^m \nu_i' g^i(x_i) + \sum_{i=0}^{m-1} \eta_i' h(u_i) \end{aligned} \quad (22)$$

has a stationary point at  $z = z^*$ .

Similarly, the corollary to Theorem 1A now becomes:

*Corollary:* At a relative minimum  $z^*$ , we have

$$\Phi(z^*) = \phi(z^*) = \sum_{i=0}^{m-1} \sigma(x_i^*, u_i^*).$$

Furthermore, the vectors  $\lambda_i$ ,  $\nu_i$ , and  $\eta_i$  must satisfy (14), (15), and

$$\lambda_{i+1} - \lambda_i = -f'_x(x_i^*, u_i^*)\lambda_{i+1} + \sigma'_x(x_i^*, u_i^*) + g^{i'}_x(x_i^*)\nu_i, \\ i = 1, \dots, m-1. \quad (23)$$

Finally, the extension of Theorem 2A gives:

*Theorem 2B:* Let  $\sigma(x, u)$  be convex and  $f(x, u)$  be linear on  $E^n \times U$ . If  $z^*$  is an admissible point and there exist vectors  $\lambda$  and nonnegative vectors  $\nu_i$  and  $\eta_i$  such that (20), (21), (23), (14), and (15) are satisfied, then  $z^*$  is a global minimum.

#### CONVEX PROGRAMMING SOLUTION

We are now in a position to consider the computational solution of the discrete optimal control problem with state constraints. We limit our discussion here to problems for which the optimality conditions are sufficient, namely,  $\sigma(x, u)$  convex and  $f(x, u)$  linear. In the interest of simplicity, we will also assume that the constraint sets  $U$  and  $X_i$  are defined by linear inequalities, that is, that the functions  $h(u)$  and  $g^i(x)$  are linear. The method for  $f(x, u)$  linear discussed here is the basis for a convergent iterative procedure for solving the more general case where  $f(x, u)$  is convex. This more general case is described in another paper (Rosen, 1966).

The general, variable coefficient linear case will be considered, that is, a discrete approximation to the differential equation

$$\dot{x} = A(t)x + B(t)u, \quad t \in [0, T]. \quad (24)$$

We will let  $A_i = A(t_i)$  and  $B_i = B(t_i)$  ( $i = 0, \dots, m$ ) and use the finite difference approximation

$$x_{i+1} - x_i = \Delta t[\theta A_{i+1}x_{i+1} + (1 - \theta)A_i x_i] + \Delta t B_i u_i, \\ i = 0, 1, \dots, m-1, \quad (25)$$

where  $0 \leq \theta \leq 1$ . For  $\theta = 0$ , this gives the explicit (forward) scheme (2), while for  $\theta = 1$  it gives the fully implicit (backward) scheme. The value  $\theta = \frac{1}{2}$  gives a numerically stable method with minimum truncation error.

The relation (25) may be solved for  $x_{i+1}$  to give

$$x_{i+1} = K_i x_i + \bar{B}_i u_i, \quad i = 0, 1, \dots, m-1, \quad (26)$$

where

$$K_i = \left[ I - \frac{\theta T}{m} A_{i+1} \right]^{-1} \left[ I + \frac{(1 - \theta)T}{m} A_i \right] \quad (27)$$

and

$$\bar{B}_i = \frac{T}{m} \left[ I - \frac{\theta}{m} A_{i+1} \right]^{-1} B_i. \quad (28)$$

The solution to the finite difference equation (26) is given by

$$x_i = Y_i x_0 + Y_i \sum_{j=0}^{i-1} \Lambda'_{i+1} \bar{B}_j u_j, \quad i = 1, \dots, m, \quad (29)$$

where the matrices  $Y_i$  satisfy the homogeneous equation

$$Y_{i+1} = K_i Y_i, \quad Y_0 = I, \quad i = 0, 1, \dots, m-1, \quad (30)$$

and where the matrices  $\Lambda_i$  satisfy the homogeneous adjoint equation

$$\Lambda_i = K'_i \Lambda_{i+1}, \quad Y_m \Lambda'_m = I, \quad i = m-1, \dots, 1. \quad (31)$$

It follows from (30) and (31) that  $Y_i \Lambda'_i = Y_{i+1} \Lambda'_{i+1} = I$ , so that  $\Lambda'_i = Y_i^{-1}$  ( $i = 1, \dots, m$ ). Furthermore, the actual calculation of  $Y_i x_0$  and of the coefficients of the  $u_i$  in (29) requires only the inversion of an  $n \times n$  matrix to get each  $K_i$  and the multiplication of  $n \times r$  matrices. These quantities are therefore readily calculated from the specified values of  $x_0$ ,  $A(t)$ ,  $B(t)$ ,  $m$ , and  $\theta$ .

Because of the linearity of (29), we can use these relations to map the original problem into the control space, that is, the product space of the  $u_i$ . This reduces the original problem to one of minimizing a convex function subject to linear inequality constraints in the space  $E^{mr}$ . Since the original problem involved  $s = m(n + r)$  variables, and since  $r \leq n$  (often with  $r = 1$  or  $r = 2$ ), this may effect a considerable reduction in the number of variables. To accomplish this reduction, we replace each  $x_i$  in the sum (1) and in the linear inequalities  $g^i(x_i) \leq 0$  which define the  $X_i$  by the corresponding righthand side of (31). Each vector  $g^i(x_i)$  thus gives rise to  $k_i$  linear inequalities on the  $u_i$ . We also have the original set of  $l$  linear inequalities  $h(u_i) \leq 0$ , which ensures that each  $u_i \in U$ . We therefore have a system of

$$ml + \sum_{i=1}^m k_i = m(l + \bar{k})$$

linear inequalities which must be satisfied by any admissible set of vectors  $u_i$ . Because of the way in which these inequalities arise, they have a special structure which can be used to advantage. We will represent the inequalities obtained from the  $g^i(x_i)$  by

$$\sum_{i=0}^{m-1} D'_i u_i - p \leq 0, \quad (32)$$

where each  $D_i$  is an  $r \times mk$  matrix and where  $p \in E^r$ . Because  $x_{i+1}$  involves only values of  $u_j$  for  $j \leq i$ , the matrix  $D' = [D'_0 D'_1 \cdots D'_{m-1}]$  has a lower triangular structure. The matrices  $D_i$  will depend on the matrices  $A_i$ ,  $B_i$ , and the matrices which define the linear transformations  $g'(x_i)$ , as well as on  $\theta$  and  $m$ . The vector  $p$  will also depend on  $x_0$ , as well as on these other quantities. The important point, however, is that the matrices  $D_i$  and the vector  $p$  can be explicitly computed with a reasonable amount of computation.

In order to simplify the discussion, we will denote by  $w \in E^{mr}$  a vector which specifies the control for  $i = 0, 1, \dots, m-1$ , that is,  $w' = (u'_0, u'_1, \dots, u'_{m-1})$ . Two subsets of  $E^{mr}$  are then given by

$$W_1 = \left\{ w \mid \sum_{i=0}^{m-1} D'_i u_i - p \leq 0 \right\} \quad (33)$$

and

$$W_2 = \{ w \mid h(u_i) \leq 0, \quad j = 0, \dots, m-1 \}. \quad (34)$$

Since it is determined by linear inequalities,  $W_1$  is closed and convex if it is not empty. Since  $W_2$  is the direct product of compact convex sets, it is compact and convex. Then

$$W = W_1 \cap W_2 \quad (35)$$

is compact and convex if it is not empty.

The first important question about the discrete problem can now be answered: Does there exist any admissible control? This is equivalent to the question: Is  $W$  an empty set? Good computational methods are available for determining if a solution to a system of linear inequalities exists and, if so, for finding such a solution. Since the inequalities of (33) and (34) have the natural form of constraints for a dual linear programming problem, a dual simplex procedure can be used for this purpose (Dantzig, 1963). The starting procedure for the gradient projection method (Rosen, 1960) is equivalent to this and may conveniently be used for this purpose. Another approach would be to use the duality theory of linear programming and to consider the primal problem corresponding to the dual constraints (33) and (34) and an arbitrary linear dual objective function. This objective function can always be chosen so as to give an initial primal feasible solution. The duality theory then says that if the primal problem has a finite maximum, the corresponding dual solution is dual feasible (that is, an admissible control). Any suitable linear programming code can therefore be used for this purpose.

Once we have determined that an admissible control exists, and in fact have actually determined such a control, we can proceed to find an optimal control. We do this by once again using (29) to eliminate the  $x_i$ , this time in the sum (1), to get a function  $\rho(w)$  to be minimized. Since convexity is preserved by a linear transformation, the function  $\rho(w)$  is convex. We have now reduced the original discrete problem to that of finding

$$\rho(w^*) = \min_{w \in W} \rho(w),$$

that is, the minimization of a convex function subject to linear inequality constraints. Furthermore, we have an admissible control  $w^0$  (determined as discussed above) with which to start the minimization procedure. A number of computationally tested methods are available for the solution of such convex nonlinear programming problems (Rosen, 1960, and Hadley, 1964). In the special case where  $\sigma(x, u)$  is linear on  $E^m \times U$ , the problem can be solved in the dual form by a dual simplex method or, in its primal form, by any primal simplex code. The possibility of formulating a discrete linear optimal control problem as a linear programming problem has been considered by Zadeh and Whalen (1962). An efficient method of solution for linear problems with large values of  $m$  has been proposed by Dantzig (1966), based on his generalized upper-bounding technique.

Once the optimal control  $w^{**} = (u_0^{**}, u_1^{**}, \dots, u_{m-1}^{**})$  has been calculated in this way, the optimal state vectors  $x_i^*$  ( $i = 1, \dots, m$ ) are immediately given by (29). The Lagrange multipliers (or shadow price vectors)  $\nu_i$  ( $i = 1, \dots, m$ ) and  $\eta_i$  ( $i = 0, \dots, m-1$ ), corresponding to the state and control constraints, are also available as part of the convex programming solution. These quantities may be of considerable interest since they give the rate of decrease in the function value with relaxation of each constraint. The influence of parameter changes on the optimal solution can also be obtained by use of the parametric solution features of many codes. Finally, if desired, the optimal adjoint vectors satisfying (23) with  $f(x, u)$  linear can be calculated from

$$\lambda_i = K(\lambda_{i+1} - \sigma'_i(x_i^*, u_i^*) - g'_i(x_i^*)\nu_i), \quad i = m-1, \dots, 1,$$

starting with  $\lambda_m = -g'_m(x_m^*)\nu_m$ .

#### COMPUTATIONAL EXAMPLE

The previous discussion will now be illustrated by means of a variable coefficient linear problem with four state variables and a scalar control.

In addition to bounded control, we also impose state constraints on one of the state variables. The system considered is in the form (24), with

$$TA(t) = tA_1 + (T - t)A_0$$

and

$$A_1 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -4 & -10 & -10 & -5 \end{bmatrix}, \quad A_0 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -8 & -16 & -12 & -5 \end{bmatrix},$$

$$B(t) = b = \begin{bmatrix} 0 \\ 0 \\ 1.0 \\ -4.5 \end{bmatrix}, \quad x_0 = \begin{bmatrix} 0.5 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

The control must satisfy  $|u(t)| \leq 1$ , so that  $h(u) = \begin{pmatrix} u - 1 \\ -u - 1 \end{pmatrix}$ . We also impose the terminal constraints  $x_1(T) = x_2(T) = 0$  and the state constraint  $|x_4(t)| \leq 0.5$  for  $0 \leq t < T$ . These give

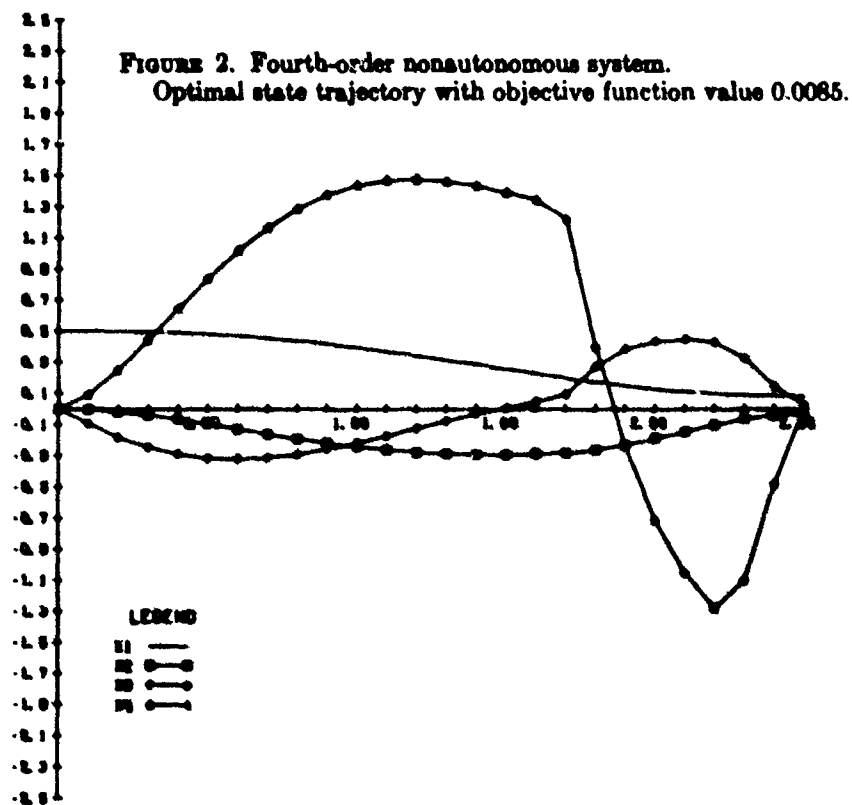
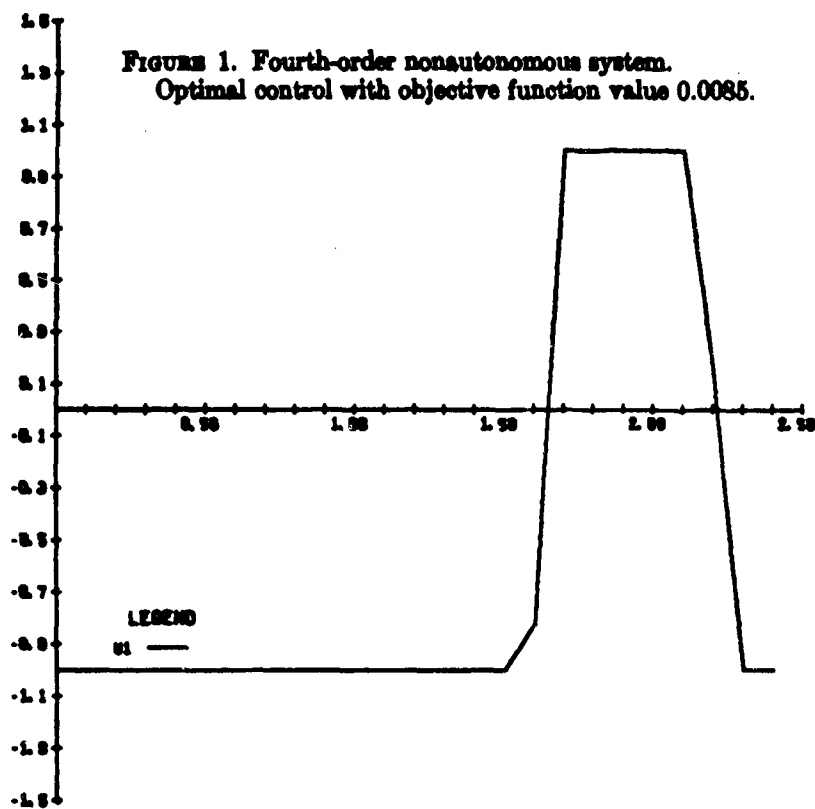
$$g^i(x(t_i)) = \begin{bmatrix} x_4(t_i) - 0.5 \\ -x_4(t_i) - 0.5 \end{bmatrix}, \quad i = 1, \dots, m-1,$$

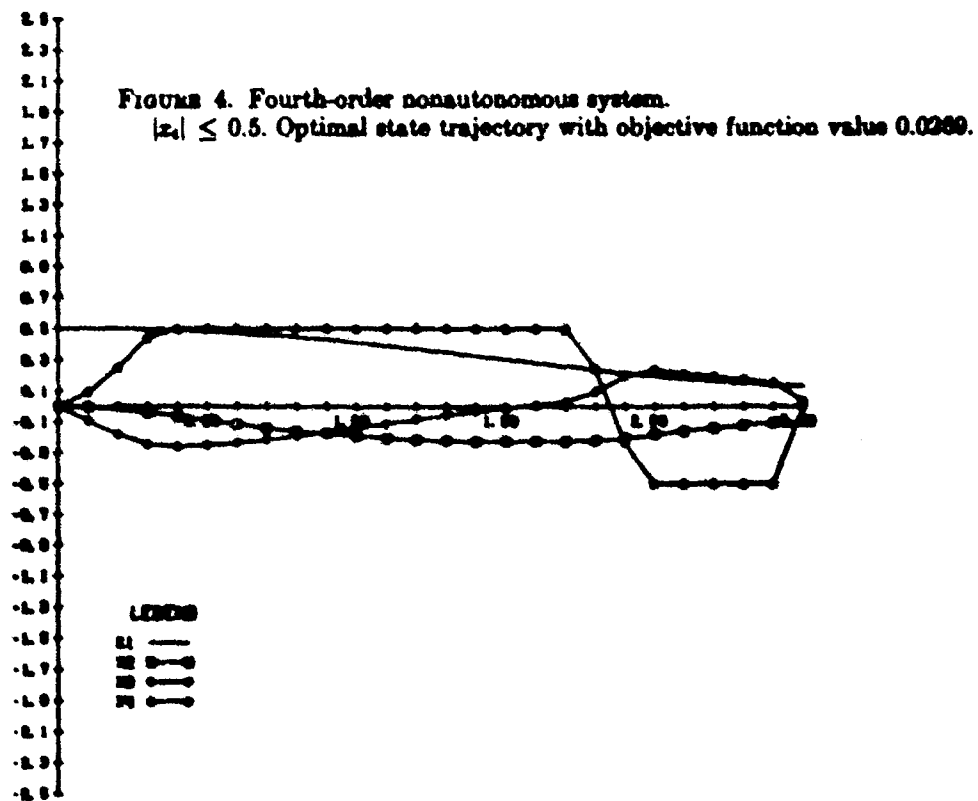
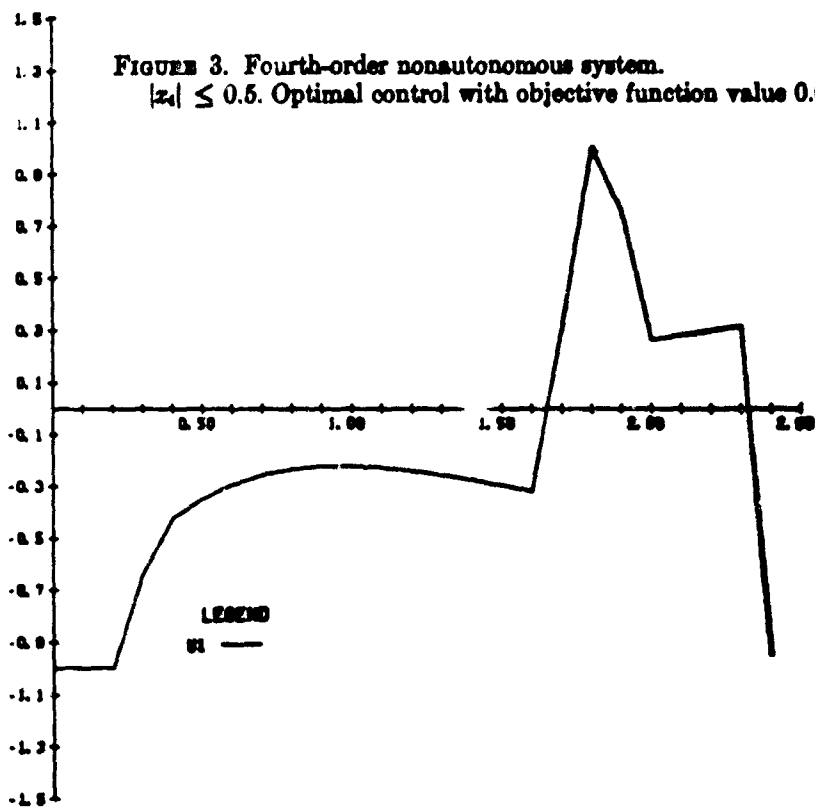
$$g^m(x(T)) = \begin{bmatrix} x_1(T) \\ -x_2(T) \\ x_4(T) \\ -x_4(T) \end{bmatrix}.$$

We wish to minimize the terminal Euclidean norm  $\|x(T)\|$ . This system is similar to a constant coefficient system for which a numerical solution has previously been obtained (Ho and Brentani, 1963).

The optimum solution to this problem, using the finite difference scheme (25) with  $\theta = \frac{1}{2}$ , is shown in Figures 1 and 2. The values  $T = 2.5$  and  $m = 25$  were used so that  $\Delta t = 0.1$ . The optimum control is shown in Figure 1, and the trajectory as given by the four state variables  $x_j(t_i)$  ( $j = 1, \dots, 4; i = 0, \dots, 25$ ) is shown in Figure 2 for the case with no state constraint on  $x_4(t_i)$ . The minimum value of the objective function attained is  $\|x(T)\|^2 = 0.008456$ . The optimal solution to the same problem with the state constraint  $|x_4(t_i)| \leq 0.5$  is shown in Figures 3 and 4. The distinct change in the control required to satisfy the state bound should be noted, as well as the increase in the terminal norm squared to 0.026922, which is due to the fact that the admissible control set  $W$  is smaller because of the state bound.

[Figures 1 through 4 appear on pages 234-6; text resumes on page 236]





These solutions were obtained by using a program based on the scheme described by (25) through (34). The convex programming problem obtained in this way was solved by using the gradient projection computer program (SHARE distribution #1399). The solution time required for each of these problems on the IBM 7090 was approximately two minutes. The program and its use to obtain the optimal solution to a variety of typical problems will be described elsewhere (Rosen and O'Hagan, 1966).

## REFERENCES

- BELLMAN, R. E., I. GLICKSBERG, and O. A. GROSS. 1958. Some aspects of the mathematical theory of control processes. R-313. Santa Monica, Calif.: The RAND Corp.
- BERGE, C. 1963. Topological spaces. New York: Macmillan, chap. 8.
- BERKOVITS, L. D. 1962. On control problems with bounded state variables. *J. Math. Anal. Appl.*, 5:488-98.
- DANTZIG, G. 1963. Linear programming and extensions. Princeton, N. J.: Princeton Univ. Press.
- . 1966. Linear control processes and mathematical programming. *J. SIAM Control Ser. A*, 4:56-60.
- HADLEY, G. 1964. Nonlinear and dynamic programming. Reading, Mass.: Addison-Wesley.
- HO, Y. C., and P. B. BRENTANI. 1963. On computing optimal control with inequality constraints. *J. SIAM Control Ser. A*, 1:319-48.
- KARLIN, S. 1959. Mathematical methods and theory in games, programming, and economics, vol. I. Reading, Mass.: Addison-Wesley.
- KUHN, H. W., and A. W. TUCKER. 1951. Nonlinear programming in Proceedings of the second Berkeley symposium on mathematical statistics and probability. Berkeley: Univ. of California Press, 481-92.
- LEITMANN, G. (ed.). 1962. Optimization techniques with applications to aerospace systems. New York: Academic Press.
- PONTRYAGIN, L. S., V. G. BOLTYANSKII, R. V. GAMKRELIDZE, and E. F. MISHCENKO. 1962. The mathematical theory of optimal processes, tr. K. N. TRIMBOFF. New York: Interscience.
- ROSEN, J. B. 1960. The gradient projection method for nonlinear programming, part I. *J. SIAM*, 8:181-217.
- . 1964. Sufficient conditions for optimal control of convex processes. Technical report CS7. Stanford, Calif.: Stanford Univ.
- . 1966. Iterative solution of nonlinear optimal control problems. *J. SIAM Control Ser. A*, 4:233-44.
- ROSEN, J. B., and M. O'HAGAN. 1966. Computational solution of constrained optimal control problems. Technical report (in preparation).
- URAWA, H. 1964. Optimal growth in a two-sector model of capital accumulation. *Rev. Econ. Studies*, 31:1-24.
- ZADEH, L. A., and B. H. WHELEN. 1962. On optimal control and linear programming. *IEEE Trans. Automatic Control*, 7:45-46.

## DISCUSSION

H. HALKIN: At the beginning of your paper, you said that the condition was necessary and sufficient. At that time you did not specify exactly the case where this would be true.

J. B. ROSEN: The question is: When are the Kuhn-Tucker conditions sufficient for the discrete control problem? The answer is that in general

they are only sufficient for a linear system of difference equations, that is, when  $f(x, u)$  in (2) is linear. This is because for sufficiency the admissible space, over which the minimization is carried out, must be a convex set. The admissible set consists of those points  $z \in E^n$ ,  $z' = (x'_1, \dots, x'_m, u'_0, \dots, u'_{m-1})$ , which satisfy (2),  $x_m \in X_m$ , and  $u_i \in U$  ( $i = 0, \dots, m-1$ ). This cannot be convex if  $f(x, u)$  is nonlinear.

J. MOSER: Is there any analysis which tells you or gives you an indication whether the discrete problem really approximates the continuous problem?

J. B. ROSEN: The question of convergence of the discrete problem optimal solution to an optimal solution of the corresponding continuous problem is closely related to the set of reachable points  $x_m \in E^n$  given by (29) with  $u_i \in U$ . For specified values of  $x_0$ ,  $A(t)$ , and  $B(t)$ , the reachable set will depend on  $m$  and  $\theta$ . What we would like is that the reachable set expands as  $\Delta t \rightarrow 0$  or  $m \rightarrow \infty$ . It can be shown that this will, in fact, be the case under certain conditions if  $\theta$  is correctly chosen in (25). On the other hand, simple examples can be constructed by using the explicit scheme ( $\theta = 0$ ), where the reachable set shrinks as  $m \rightarrow \infty$ . In such a situation, one may find that, as the grid size is decreased ( $m \rightarrow \infty$ ), a value exists such that for all larger  $m$  it is no longer possible to reach the terminal manifold. In such a case, one clearly does not have convergence.

J. MOSER: I think it would be an interesting question to investigate the conditions which would ensure convergence.

H. HALKIN: There is a paper by Professor Markus (paper 6) on the stability of solutions of optimal control problems with respect to changes in the data of the problems. I think that if a problem is stable in Professor Markus' sense, then the solution of a discretisation of this problem will tend to the solution of the problem itself as the discretisation is made finer. Professor Markus gives an answer to such a problem in the case of a linear system with constant coefficients.

Reprinted from  
J. SLAM Control  
Vol. 4, No. 1, 1966  
Printed in U.S.A.

## ITERATIVE SOLUTION OF NONLINEAR OPTIMAL CONTROL PROBLEMS\*

J. B. ROSEN†

**Abstract.** The solution of nonlinear, state-constrained, discrete optimal control problems by mathematical programming methods is described. The iterative solution consists essentially of Newton's method with a convex (or linear) programming problem solved at each iteration. Global convergence of the iterative method is demonstrated provided a convexity and constraint set condition are both satisfied. The computational solution of nonlinear equation control problems makes use of a previously developed method for state-constrained linear equation problems. The solution method for nonlinear problems is illustrated by means of two numerical examples.

**1. Introduction.** The optimal control problem considered here is a rather general type of discrete problem. We wish to minimize a convex function of the state and control vectors, where the control vectors must lie in a specified convex set. In addition the state vectors must also satisfy specified constraints at each discrete time, as well as initial and terminal conditions. Furthermore, the system dynamics may be given by a nonlinear recursion relation provided that the nonlinearity is convex in an appropriate way. A discrete system of the type considered here may represent a process which is actually discrete (see, for example, [3], [1]), or it may be obtained from a finite difference approximation to a continuous system in which we wish to minimize a convex functional. Such an approximation is *always* required when a numerical integration, using a digital computer, is part of the solution process.

The purpose of this report is to describe a computational method for solving this general type of discrete problem, and to show by means of the relevant theorems that the method will always work when the appropriate assumptions are satisfied. The method is an iterative procedure that determines a sequence of admissible trajectories (state and control vectors satisfying all constraints); the sequence converging to an admissible trajectory that satisfies the necessary conditions for optimality. The method has been used to obtain numerical solutions to several small nonlinear test problems. In addition to showing that it is not difficult to implement the

\* Received by the editors June 28, 1965. Presented at the First International Conference on Programming and Control, held at the United States Air Force Academy, Colorado, April 15, 1965.

† Computer Sciences Department and Mathematics Research Center, University of Wisconsin, Madison, Wisconsin. This research was sponsored in part by the National Aeronautics and Space Administration under Research Grant NGR-50-002-028 and in part by the Mathematics Research Center under Contract No. DA-11-022-ORD-2059.

scheme described here, these numerical results show that, at least for the test problems considered, the number of iterations required is small.

In a previous publication [14] a statement of the Kuhn-Tucker conditions was given for the nonlinear state-constrained problem considered here. A computational procedure for systems described by linear recursion relations was also given based on a convex (or linear) programming computer code. Numerical results described there show that this computational procedure is efficient for typical linear systems. The method described in the present paper takes advantage of this efficiency by solving a sequence of such linear problems. From this point of view the method of the present report may be thought of as Newton's method (see, for example, [9]) with a convex (or linear) programming problem solved at each iteration. The use of various forms of Newton's method for the numerical solution of optimal control problems has been proposed in a number of earlier publications [4], [6], [10], [12]. The two important differences between the method described here and these earlier proposals are that (1) in the present method *global* convergence is assured when a convexity and constraint set condition are both satisfied, and (2) large changes in both the control and state vectors may take place at each iteration until these vectors are close to their limiting values, thereby greatly accelerating convergence during the early states. The limiting convergence rate is quadratic, as expected in Newton's method.

Another way of looking at this method for nonlinear problems is that at each iteration we get an admissible and optimal trajectory which satisfies a linear recursion relation which differs to some extent from the true nonlinear recursion relation. At each iteration the amount by which the linearization is in error decreases, so that in the limit the trajectory obtained is an *optimal* solution to the *linearized* problem obtained by linearizing about the limiting trajectory. Since it is the recursion relation which is linearized, the limiting trajectory is the optimal solution to a control problem described by linear recursion relations. It therefore follows that for the class of discrete nonlinear problems considered, the optimal solution has the properties of a solution to a discrete problem with linear recursion relations.

The requirement that the state vectors satisfy specified constraints usually increases the difficulty of the optimal control problem (see, for example, [5] and [13, Chap. 6]). In the approach used here to solve the state-constrained discrete problem, the convergence proof uses the fact that the state vector at each discrete time belongs to a convex compact set. In this sense then, the liability of the state-constrained problem has now become an asset. The existence of state constraints also introduces a symmetry into the problem, so that the usual sharp distinction between the

(independent) control vectors and (dependent) state vectors largely disappears.

The method described here applies to a recursion relation in the form of a system of inequalities, and might represent a finite difference approximation to a system of differential *inequalities*. By the use of a modified objective function, the problem usually considered corresponding to a system of differential equations can be handled. The "classical" two-point boundary value problem can also be solved in this fashion by allowing the control vector to represent the error in the difference equations and minimizing this error.

It should be emphasized that while the convexity assumption is needed in order to prove convergence, the computational method can be applied even when this assumption is not satisfied. In many such cases the iterative method will still converge, and if so, the trajectory obtained will satisfy the necessary conditions for an optimal trajectory. Furthermore, at each iteration a linear constraint minimization problem with either a convex or linear function is solved. Because of this, the method will almost always converge to a trajectory, which is at least a local minimum of the objective function, rather than an arbitrary stationary trajectory. It should also be mentioned that the method considered here requires only the Jacobian matrix (first partial derivatives) of the system equations, and does not need the Hessian matrix (second partial derivatives) as required by some other computational schemes [6], [10], [12]. For many nonlinear problems this may permit a great reduction in the computation required.

While the iterative method described was developed for problems arising in control theory, it may also be used to solve any finite-dimensional constrained minimization problem of the general type considered. In this respect the method is also a contribution to the solution of nonconvex mathematical programming problems.

**2. Problem formulation.** The discrete optimal control problem we shall consider here is to determine  $m + 1$  state vectors  $x_i^* \in E^n$  and  $m$  control vectors  $u_i^* \in E^r$  which satisfy (2.2), (2.3) and (2.4) and such that

$$(2.1) \quad \sum_{i=0}^{m-1} \sigma(x_i^*, u_i^*) = \min \sum_{i=0}^{m-1} \sigma(x_i, u_i)$$

for all vectors  $x_i$  and  $u_i$  that satisfy the recursion relation

$$(2.2) \quad x_{i+1} - x_i = f(x_i, u_i), \quad i = 0, 1, \dots, m-1,$$

with

$$(2.3) \quad u_i \in U_i \subset E^r, \quad i = 0, 1, \dots, m-1,$$

and

$$(2.4) \quad x_i \in X_i \subset E^n, \quad i = 0, 1, \dots, m.$$

The subsets  $X_i$  and  $U_i$  are assumed to be compact and convex. We assume that  $\sigma$  is a convex function from each direct product  $X_i \times U_i$  to  $E^1$ . We also assume that  $f$  is a function from each  $X_i \times U_i$  to  $E^n$ . An additional assumption on the differentiability and convexity of the components of  $f$  will be needed later. It should be mentioned that the results obtained actually hold (with obvious modification) for the more general case where  $\sigma$  and  $f$  may depend explicitly on the index  $i$ . When the discrete problem is obtained from a continuous problem, this corresponds to the explicit dependence of  $\sigma$  and  $f$  on time. However, in order to avoid the complication of additional subscripts we will limit consideration to the simpler case.

A discrete problem of this type may arise directly, or it may arise as a finite difference approximation to a continuous system. For example, suppose that in the original continuous system we wish to determine a control  $u(t)$  with range  $U(t)$  for each  $t \in [0, T]$ , and a trajectory  $x(t)$  with range  $X(t)$  for each  $t \in [0, T]$ , such that the functional

$$(2.5) \quad \int_0^T \sigma(x(t), u(t)) dt$$

is minimised, and  $x(t)$  and  $u(t)$  satisfy the system of differential equations

$$(2.6) \quad \dot{x} = \bar{f}(x, u), \quad t \in [0, T].$$

The sum (2.1) then represents the simplest approximation to the integral (2.5), and the recursion relation (2.2) the simplest finite difference approximation to the system (2.6), if we let  $\Delta t = T/m$ ,  $\sigma = \Delta t \bar{\sigma}$ , and  $f = \Delta t \bar{f}$ . The form of (2.2) may be retained even when more sophisticated finite difference schemes are used to approximate (2.6), but the relationship between  $f$  and  $\bar{f}$  will become more complicated. The use of a more accurate implicit finite difference scheme when  $f$  is linear has been considered in [14]. It should be emphasized that in this paper we solve the discrete problem for a fixed value of  $m$ , and that we are interested in convergence (for fixed  $m$ ) to an exact solution of the nonlinear discrete problem. The convergence to the solution of the continuous problem as  $m \rightarrow \infty$  will not be considered here.

In order to show convergence of the iterative procedure we will consider the discrete system (2.1), (2.3) and (2.4), with (2.2) replaced by the system of inequalities

$$(2.7) \quad x_{i+1} - x_i \leq f(x_i, u_i), \quad i = 0, 1, \dots, m-1.$$

Such a system of inequalities may arise as a discrete approximation to a

system of differential inequalities of the form  $\dot{x} \leq f(x, u)$ . On the other hand, if one really wants to solve (2.2), this is accomplished by obtaining an optimum solution to (2.7) with an appropriately modified objective function, as discussed at the end of this section.

In order to simplify notation we proceed as in [14], and denote a specific control  $(u_0', u_1', \dots, u_{m-1}')$  and corresponding trajectory  $(x_0', x_1', \dots, x_m')$  by a single vector  $z \in E^s$ , where  $s = m(r + n) + n$ . Thus, a solution to the discrete system is specified by the vector

$$(2.8) \quad z' = (x_0', x_1', \dots, x_m', u_0', u_1', \dots, u_{m-1}').$$

We will also denote by  $Z \subset E^s$  the direct product of the sets  $X_i$  and  $U_i$ , so that

$$(2.9) \quad Z = \prod_{i=0}^m X_i \times \prod_{i=0}^{m-1} U_i.$$

Since the sets  $X_i$  and  $U_i$  are convex and compact,  $Z$  is also convex and compact. We can now represent the objective by means of the function

$$(2.10) \quad \phi(z) = \sum_{i=0}^{m-1} \sigma(x_i, u_i).$$

It follows from our assumption concerning  $\sigma$  that  $\phi(z)$  is convex on  $Z$ . Finally we represent the  $l = mn$  equations (2.2) or inequalities (2.7) by means of a function  $v(z)$  from  $E^s$  to  $E^l$ . We let

$$(2.11) \quad v_{i,j} = f_j(x_i, u_i) + x_{i,j} - x_{i+1,j}, \\ i = 0, 1, \dots, m-1, \quad j = 1, \dots, n.$$

The equations (2.2) are then given by  $v(z) = 0$ , and the inequalities (2.7) by  $v(z) \geq 0$ . In this notation we can restate our problem (2.1), (2.3), (2.4) and (2.7) as follows:

$$(2.12) \quad \phi(z^*) = \min_z \{ \phi(z) \mid z \in Z, v(z) \geq 0 \}.$$

Some remarks on the nature of the admissible set

$$S = \{ z \mid z \in Z, v(z) \geq 0 \}$$

are in order here. The set  $Z$  is by assumption convex and compact, and in fact will usually be a polyhedral set in  $E^s$ . The admissible set corresponding to the original discrete problem (2.2), (2.3) and (2.4) is given by

$$S_1 = \{ z \mid z \in Z, v(z) = 0 \}.$$

The set  $S_1$  is convex only if  $v(z)$  is linear in  $z$ , that is,  $f(x, u)$  is linear in  $x$  and  $u$ . If one or more components of  $f$  are nonlinear in  $x$  or  $u$ , the set  $S_1$  is

nonconvex. For a general nonlinear function  $f(x, u)$ , the set  $S$  is also nonconvex. The iterative procedure of the following sections can be applied to such problems and will, in fact, often converge. However, there is no guarantee in the case of a general nonlinear  $f$  that the procedure will always converge. In order to prove convergence we require that each component of  $v(z)$  be a *convex* function. It should be emphasized that this is *not* the requirement which makes  $S$  a convex set (except in the limiting case where  $v(z)$  is linear). The set  $S$  is convex if each component of  $v(z)$  is a *concave* function. Thus the convergence argument holds for the minimization of a convex function over a certain kind of nonconvex region.

If we actually want to satisfy (2.2) we must obtain a solution to the problem  $\phi(z^*) = \min_{z \in S_1} \phi(z)$ ; that is, we require  $v(z^*) = 0$ . In order to achieve this and still solve a problem in the form of (2.12) we let

$$(2.13) \quad \bar{\phi}(z) = \phi(z) + \alpha \sum_{i,j} v_{i,j},$$

where  $\alpha$  is a sufficiently large positive constant. Since each component  $v_{i,j}$  is a convex function,  $\bar{\phi}(z)$  is a convex function. We then solve  $\min_{z \in S} \bar{\phi}(z)$ , which is in the form of (2.12). It is shown in the Appendix that provided the constraint set  $S$  satisfies a certain condition (essentially the same condition which insures convergence) there will always exist a value of  $\alpha$  such that any local minimum of  $\bar{\phi}(z)$  for  $z \in S$  is also a local minimum of  $\phi(z)$  for  $z \in S_1$ .

We are now able to describe the iterative method for solving the discrete optimal control problem in terms of the (in general, nonconvex) mathematical programming problem (2.12).

**3. Linearized problem.** Let  $Z$  be a compact convex subset of  $E^*$ , and  $v(z)$  be a function from  $Z$  to  $E'$  with  $v \in C^1(Z)$ . We assume that for some  $z^0 \in Z$  we have  $v(z^0) > 0$  and define a subset of  $E^*$  by

$$(3.1) \quad S = \{z \mid z \in Z, v(z) \geq 0\}.$$

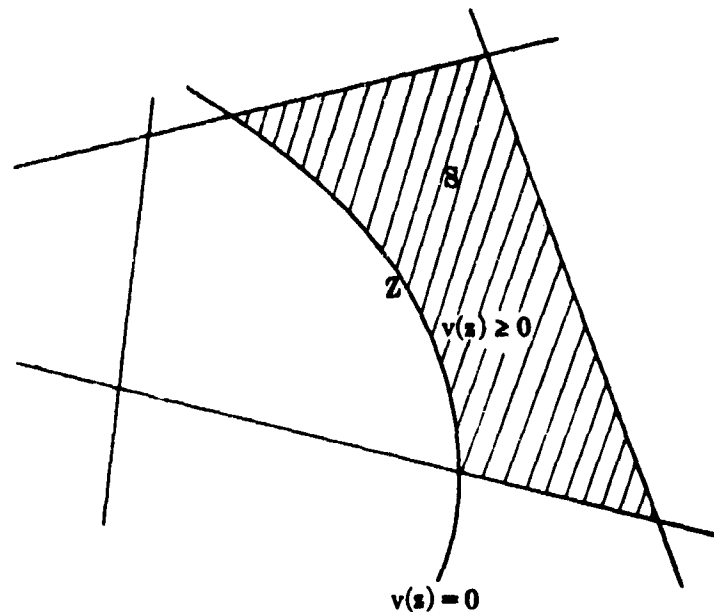
Since  $z^0 \in S$ , the set  $S$  is not empty. Also since  $S$  is a closed subset of  $Z$  it is compact but, in general, not convex (see Fig. 1).

If we let  $v_i(y)$  be the  $l \times s$  Jacobian matrix of  $v$  evaluated at  $z = y$ , we can define for each fixed  $y \in Z$  the linear function on  $Z$ ,

$$(3.2) \quad w(z, y) = v(y) + v_i(y)[z - y].$$

For each  $y \in Z$  we obtain a subset of  $E^*$  given by

$$(3.3) \quad W(y) = \{z \mid w(z, y) \geq 0\}.$$


 FIG. 1. The convex set  $Z$  and subset  $S$ 

Now we consider the point-to-set mapping

$$(3.4) \quad \Gamma : Z \rightarrow Z,$$

given by

$$(3.5) \quad \Gamma y = W(y) \cap Z.$$

This is illustrated in Fig. 2.

**THEOREM 1.** *The set  $\Gamma y$  is compact and convex. Furthermore, if each component of  $v(z)$  is convex on  $Z$ , then for each  $y \in S$ ,*

$$(3.6) \quad y \in \Gamma y \subset S.$$

*Proof.* For each  $y$ , the set  $W(y)$  is the intersection of  $l$  halfspaces, a closed convex set. Therefore the intersection of  $W(y)$  and the compact convex set  $Z$  is compact and convex. Next we note that since  $y \in S$ ,

$$(3.7) \quad w(y, y) = v(y) \geq 0,$$

so that  $y \in W(y)$ . Then since  $y \in Z$ , we have  $y \in \Gamma y$ .

Furthermore, by the convexity of  $v(z)$ , we have for any  $(y, z) \in S \times S$ ,

$$(3.8) \quad v(z) \geq v(y) + v_*(y)[z - y] = w(z, y).$$

Then for each  $z \in W(y)$ ,

$$(3.9) \quad v(z) \geq w(z, y) \geq 0,$$

so that for every  $z \in W(y) \cap Z$  we have  $z \in S$ , or  $\Gamma y \subset S$ .

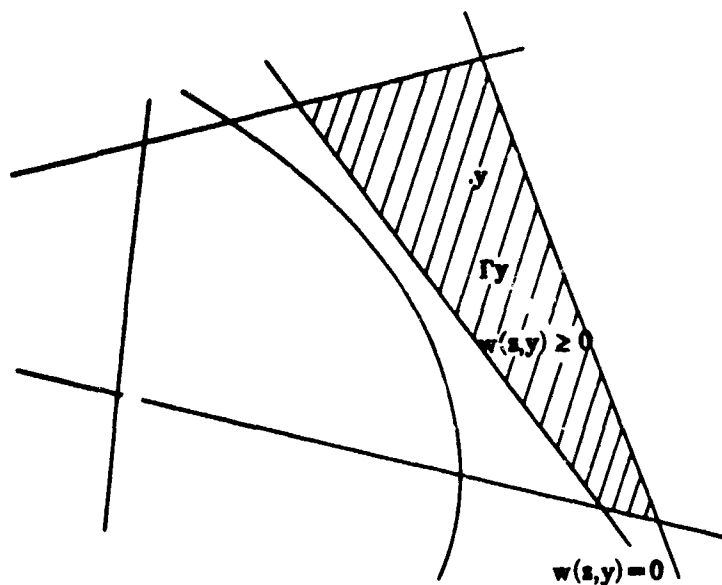


FIG. 2. The convex subset  $\Gamma y \subset S$  for  $y \in S$

Directly from (3.6) we get the following.

COROLLARY.  $\Gamma y$  maps  $S$  onto  $S$ .

The constraints for the problem have now been defined in terms of the convex subset  $Z$  and the function  $v(z)$ . The objective function is given by a function  $\phi(z)$  from  $Z$  to  $E^1$  which is continuous and convex on  $Z$ . The iterative procedure, starting with an initial point  $y^0 \in S$  can now be stated in a concise form. A sequence  $\{y^j\}$  is obtained which satisfies

$$(3.10) \quad \phi(y^{j+1}) = \min_{z \in \Gamma y^j} \phi(z), \quad j = 0, 1, \dots$$

Such a sequence is obtained by solving a well behaved convex constrained minimization problem with  $z \in \Gamma y^j$ , to get the minimum  $\phi(y^{j+1})$  at a point  $y^{j+1} \in \Gamma y^j$ . The convexity of the subset  $\Gamma y^j$  and the function  $\phi(z)$  insure that a global minimum of  $\phi(z)$  for  $z \in \Gamma y^j$  is attained at  $z = y^{j+1}$ .

Suppose that the sequence  $\{y^j\}$  converges to a limit point  $y^*$ . We would like to be able to state that the point  $y^*$  is the optimum solution to the partially linearized problem obtained by linearizing the constraints  $v(z) \geq 0$ , about  $z = y^*$ . That is, we want

$$(3.11) \quad \phi(y^*) = \min_{z \in \Gamma y^*} \phi(z).$$

In terms of the original discrete optimal control problem (2.1), (2.3), (2.4) and (2.7), this is equivalent to the statement that the control  $u_i^*$ ,  $i = 0, 1, \dots, m-1$ , and trajectory  $x_i^*$ ,  $i = 0, 1, \dots, m$ , give an

optimal solution to the problem obtained by linearising (2.7) about  $u_i^*$  and  $x_i^*$ .

However, without some further assumption, the relationship (3.11) may not hold. This is shown by the following simple two-dimensional example. Let

$$(3.12) \quad Z = \{z \mid 0 \leq z_1 \leq 1, 0 \leq z_2 \leq 1\}$$

and

$$(3.13) \quad v(z) = 4(z_1 - \frac{1}{2})^2 - z_2,$$

so that the feasible set  $S$  is given by

$$(3.14) \quad S = \{z \mid 4(z_1 - \frac{1}{2})^2 - z_2 \geq 0, 0 \leq z_1 \leq 1, 0 \leq z_2 \leq 1\}.$$

This is illustrated in Fig. 3. Also let  $\phi(z) = z_1$ . We have

$$(3.15) \quad w(z, y) = v(y) + 8(y_1 - \frac{1}{2})(z_1 - y_1) - (z_2 - y_2),$$

so that for  $y^0 = (1, 0)$  we get

$$(3.16) \quad \Gamma y^0 = \{z \mid 4z_1 - z_2 - 3 \geq 0, 0 \leq z_1 \leq 1, 0 \leq z_2 \leq 1\}.$$

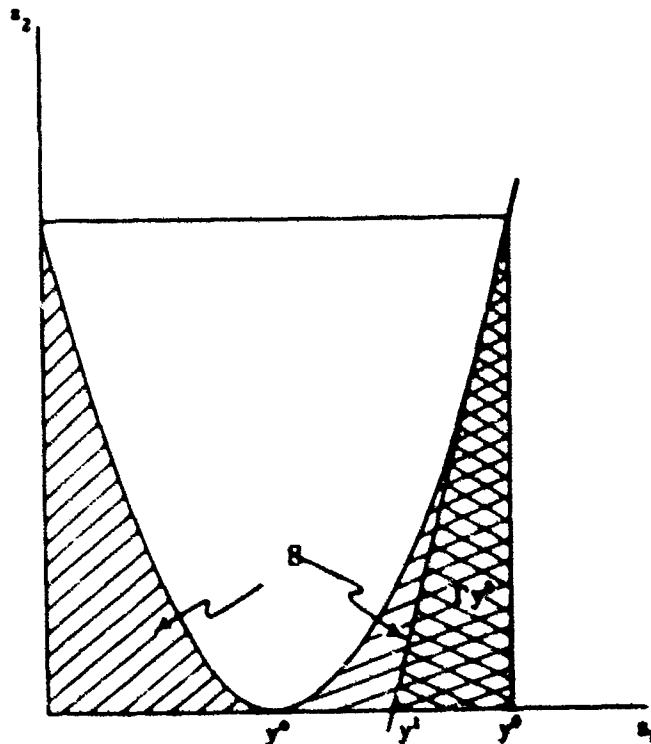


FIG. 3. Two-dimensional example

The solution to (3.10) for  $j = 0$  is easily seen (from Fig. 3) to be  $y^1 = \frac{1}{2}$ . The sequence  $\{y^j\}$  obtained in this way converges to  $y^* = (\frac{1}{2}, 0)$ , with  $\phi(y^*) = \frac{1}{2}$ . But  $\Gamma y^*$  is the interval  $[0, 1]$  on the  $z_1$  axis, so that  $\min_{z \in \Gamma y^*} \phi(z) = 0$ , and is attained at  $z = (0, 0) \neq y^*$ .

In order that the limit point  $y^*$  always satisfy (3.11) it is sufficient that the mapping  $\Gamma y$  be continuous. The mapping  $\Gamma y$  is continuous (both upper and lower semicontinuous) if for any point  $y^1 \in S$  and any point  $y^2 \in S$  in the neighborhood of  $y^1$ , there is *some* point of  $\Gamma y^1$  close to *each* point of  $\Gamma y^2$ . The continuity of  $\Gamma y$  follows from two assumptions we make concerning the set  $S$ .

(1) For each  $y \in S$ , the Jacobian matrix  $v_*(y)$  has full row rank, that is,  $\text{rank} = l \leq s$ .

(2) For each  $y \in S$ , the convex set  $\Gamma y$  contains interior points.

These two assumptions are essentially the Kuhn-Tucker constraint qualification for the set  $S$  (see, for example, [2]). The proof that (1) and (2) imply the continuity of  $\Gamma y$  is given in the Appendix. A slightly stronger assumption than (2), which however involves only the rank of an augmented Jacobian matrix, is also given there.

The difficulty in the previous two-dimensional example occurs because the assumption (2) above is not satisfied. In particular, for  $y^* = (\frac{1}{2}, 0)$ ,  $\Gamma y^*$  is just the interval  $[0, 1]$ . As a result the mapping  $\Gamma y$  is not continuous in the neighborhood of  $y^*$ .

The first assumption above is always satisfied when the function  $v(z)$  is defined by (2.11), as shown in the following.

**LEMMA.** *If  $v(z)$  corresponds to the discrete recursion relation, as given by (2.11), then assumption (1) is satisfied.*

*Proof.* Directly from (2.11) we have that

$$\begin{aligned} \frac{\partial v_{i,j}}{\partial x_{i+1,j}} &= -1, \\ (3.17) \quad \frac{\partial v_{i,j}}{\partial x_{i+1,p}} &= 0, \quad p \neq j, \\ \frac{\partial v_{i,j}}{\partial x_{q,p}} &= 0, \quad q > i+1, \quad p = 1, \dots, n, \end{aligned}$$

for  $i = 0, 1, \dots, m-1; j = 1, \dots, n$ . Therefore the Jacobian matrix  $v_*$  contains a square  $(mn \times mn)$  lower triangular matrix with elements  $-1$  along its diagonal. Since such a matrix is nonsingular and since  $v_*$  has  $mn$  rows,  $v_*$  has full row rank.

**4. Convergence of iterative procedure.** The iterative procedure will now be considered in more detail. We again consider the convex function  $\phi$  from

$Z$  to  $E^1$ , with  $\phi \in C^1(Z)$ . Since  $S$  is compact,  $\phi(z)$  is bounded and attains its minimum for  $z \in S$ . In particular, let

$$(4.1) \quad \mu = \min_{z \in S} \phi(z).$$

For each  $y \in S$ , the set  $\Gamma y$  is compact so that the minimum of  $\phi(z)$  for  $z \in \Gamma y$  is attained. We let

$$(4.2) \quad \Psi(y) = \min_{z \in \Gamma y} \phi(z).$$

We now show that because of the continuity of  $\Gamma y$ , the function  $\Psi(y)$  is continuous for  $y \in S$ .

LEMMA.  $\Psi(y)$  is continuous for  $y \in S$ .

*Proof.* For  $y^1 \in S$ , let  $\Psi(y^1)$  be attained at  $z^1 \in \Gamma y^1$ , that is  $\Psi(y^1) = \phi(z^1)$ . Now choose  $y^2 \in S$  close to  $y^1$ , and let  $\Psi(y^2)$  be attained at  $z^2$ , so that  $\Psi(y^2) = \phi(z^2)$ . Suppose  $\phi(z^2) \leq \phi(z^1)$ . Now by the continuity of  $\Gamma y$  we can choose  $\tilde{z}^1 \in \Gamma y^1$  close to  $z^2$ . Then by the continuity of  $\phi(z)$  we have  $\phi(\tilde{z}^1)$  close to  $\phi(z^2)$ . But since  $\phi(z^1) \leq \phi(z)$  for every  $z \in \Gamma y^1$ , we have

$$(4.3) \quad \phi(z^2) \leq \phi(z^1) \leq \phi(\tilde{z}^1),$$

so that  $\phi(z^1)$  is close to  $\phi(z^2)$ .

A similar argument holds for  $\phi(z^1) \leq \phi(z^2)$ .

Starting with  $y^0 \in S$  we generate a sequence of vectors  $\{y^j\}$  as follows:

$$(4.4) \quad \phi(y^{j+1}) = \min_{z \in \Gamma y^j} \phi(z), \quad j = 0, 1, \dots$$

Note that if  $Z$  is a polyhedral set then  $\Gamma y^j$  is a polyhedral set determined by specified linear inequalities. Furthermore,  $\phi(z)$  is a convex function, so that for each  $y^j$  we solve a straightforward convex programming problem with linear constraints.

THEOREM 2. Every vector of the sequence  $\{y^j\}$  is in  $S$ . The corresponding sequence of values  $\{\phi(y^j)\}$  is monotonically decreasing. The sequence  $\{y^j\}$  contains a convergent subsequence converging to a point  $y^* \in S$  such that

$$(4.5) \quad \mu \leq \phi(y^*) \leq \phi(y^j), \quad j = 0, 1, \dots,$$

and

$$(4.6) \quad \phi(y^*) = \min_{z \in \Gamma y^*} \phi(z).$$

*Proof.* By Theorem 1, we have  $y^j \in \Gamma y^j \subset S$ , so that each  $y^j$  is in  $S$ . Also since  $y^j \in \Gamma y^j$  we must have

$$(4.7) \quad \phi(y^{j+1}) = \min_{z \in \Gamma y^j} \phi(z) \leq \phi(y^j),$$

so that  $\{\phi(y^j)\}$  is monotonically decreasing.

Since  $S$  is bounded the sequence  $\{y^j\}$  contains a convergent subsequence. Let  $y^*$  be the limit point of such a convergent subsequence. Since  $S$  is compact,  $y^* \in S$ , and  $\phi(y^*) \geq \mu$ . Furthermore, from the monotonicity of the sequence  $\{\phi(y^j)\}$  the relation (4.5) must hold.

To demonstrate (4.6), we observe that since  $y^* \in S$ , we have  $y^* \in \Gamma y^*$ , so that

$$(4.8) \quad \Psi(y^*) = \min_{z \in \Gamma y^*} \phi(z) \leq \phi(y^*).$$

Now suppose that  $\Psi(y^*) < \phi(y^*)$ . Then by the continuity of  $\Psi(y)$  we can pick  $k$  sufficiently large so that  $\Psi(y^k) < \phi(y^*)$ . But from (4.2) and (4.4) we have  $\phi(y^{k+1}) = \Psi(y^k)$ , so that  $\phi(y^{k+1}) < \phi(y^*)$ , which contradicts (4.5). Therefore we must have  $\Psi(y^*) = \phi(y^*)$ .

**THEOREM 3.** *Let  $y^*$  be a limit point of  $\{y^j\}$ . Then  $y^*$  is the global minimum of the partially linearized problem about the point  $y^*$ . Furthermore, the optimality conditions (the Kuhn-Tucker necessary conditions) which must be satisfied at a global minimum of the problem (2.12) are, in fact, satisfied at  $y^*$ .*

*Proof.* The set  $\Gamma y^*$  is the intersection of  $Z$  and the convex set  $W(y^*)$  obtained by linearizing the constraints  $c(z) \geq 0$ , about  $z = y^*$ . It follows immediately from (4.6) that  $y^*$  is a global optimum solution to this partially linearized problem.

As mentioned in the previous section, the assumptions (1) and (2) on the set  $S$  are equivalent to the Kuhn-Tucker constraint qualification. It is shown in their original paper [11] that with this qualification the optimum solution  $z^*$  to a general nonlinear problem has the property that the gradient  $\nabla \phi(z^*)$  must belong to the convex cone of inward normals to the active constraints at  $z^*$ . The solution  $y^*$  to the partially linearized problem about  $y^*$  will, of course, also have this property. Therefore,  $\nabla \phi(y^*)$  belongs to the convex cone of inward normals to the active constraints at  $y^*$ , i.e., the Kuhn-Tucker necessary conditions for a global minimum are satisfied at  $y^*$ .

**5. Computational solution.** The computational solution of the nonlinear discrete optimal control problem (2.1)–(2.4) is considered in this section. We will assume that the convex compact sets  $U_i$  and  $X_i$  are convex polytopes defined by specified linear inequalities (see Appendix). In order to apply the computational method we need only make the additional assumption that the functions  $\sigma(x, u)$  and  $f(x, u)$  are of class  $C^1$  on each  $X_i \times U_i$ . However, in order to insure the validity of the convergence proof (Theorem 2) we must make an additional assumption concerning  $f$  and an assumption about the linear inequalities defining the  $X_i$  and  $U_i$ . We assume that each component  $f_j$  of  $f$ ,  $j = 1, \dots, n$ , is either convex or concave on  $X_i \times U_i$ .

For  $i = 0, 1, \dots, m-1$  and  $j = 1, \dots, n$  we let

$$(5.1) \quad \begin{aligned} \bar{v}_{i,j} &= f_j(x_i, u_i) + x_{i,j} - x_{i+1,j}, \\ v_{i,j} &= \begin{cases} \bar{v}_{i,j} & \text{for } f_j \text{ convex on } X_i \times U_i, \\ -\bar{v}_{i,j} & \text{for } f_j \text{ concave on } X_i \times U_i. \end{cases} \end{aligned}$$

The function  $v(z)$ , with components  $v_{i,j}$ , is thus a convex function on  $Z$ . Furthermore, the equations (2.2) are now equivalent to  $v(z) = 0$ .

As discussed in the Appendix the linear inequalities which define the  $X_i$  and  $U_i$  are specified in terms of the vector  $z$  by  $a_i'z - b_i \geq 0$ ,  $i = 1, \dots, k$ , giving the polyhedral set  $Z$ . We make the following assumption about these linear inequalities. Let  $\bar{y} \in S$  be a boundary point of  $Z$ , i.e.,  $v(\bar{y}) = 0$ , and  $a_i'\bar{y} - b_i = 0$ ,  $i = 1, \dots, k$ . Then the  $(l+k) \times s$  matrix consisting of  $v_i(\bar{y})$  augmented by the rows  $a_i'$ ,  $i = 1, \dots, k$ , is of full row rank ( $=l+k$ ). According to the Lemma at the end of §3,  $v_i(y)$  is always of full row rank, so this assumption is essentially a condition on the vectors  $a_i$ . As shown in the Appendix it follows from the full rank condition that  $\Gamma y$  is a continuous mapping. The convergence proof of Theorem 2 is applicable because  $v(z)$  is convex and  $\Gamma y$  is continuous.

At each iteration we wish to solve a mathematical programming problem of the form,

$$(5.2) \quad \min \{ \phi(z) \mid a_i'z - b_i \geq 0, i = 1, \dots, k; w(z, y) \geq 0 \}.$$

This is a linear constraint problem with  $m(r+n) + n$  variables and  $k+l$  constraints. For small problems a direct computational solution of (5.2) causes no difficulty. In many practical cases however, the number of state variables is greater than the number of control variables, i.e.,  $r < n$ . In such a case there is a considerable computational advantage in treating the linearized problem (5.2) as the linear problem was treated in [14]. In effect, the linear relations  $w(z, y) = 0$  are used to solve explicitly for the vectors  $x_i$ ,  $i = 1, \dots, m$ , in terms of  $x_0$  and the  $u_i$ ,  $i = 0, 1, \dots, m-1$ . Substitution for the vectors  $x_i$  in  $\phi(z)$  and the inequalities  $a_i'z - b_i \geq 0$  reduces the original problem (5.2) to one in only  $mr + n$  variables. This reduced problem may then be solved by an appropriate linear constraint method which takes advantage of the particular form of  $\phi$ . For example, if  $\phi$  is quadratic, a quadratic programming method may be used.

In the important case where  $\phi$  is linear, a further efficiency is made possible by treating the reduced problem as the dual problem, and solving the corresponding primal linear programming problem. This permits us to take advantage of the fact that the variables of the reduced problem (the control variables) are not required to be nonnegative, and that there are

more inequality constraints than variables. The corresponding primal problem consists of  $mr + n$  equations in  $mn + k$  nonnegative variables. The numerical examples discussed below are of this type.

The use of the linear equality relations  $w(z, y) = 0$  has the additional computational advantage that no modification of the true objective function is required. On the other hand a possible theoretical difficulty may arise since even with  $v(z)$  convex it is usually not true that  $y' \in \Gamma y'$  when  $\Gamma y$  is determined by  $w(z, y) = 0$ . Thus the monotone behavior of  $\phi(y')$  is not guaranteed. However, no such difficulty has been observed in the actual numerical calculations.

In order to illustrate the application of the iterative method we will discuss two numerical solutions to a nonlinear problem. The problem considered is a discrete approximation to the following continuous scalar ( $n = 1$ ) problem:

$$\min \int_0^1 u(t) dt,$$

subject to  $\dot{x} = f(x, u)$ ,  $|u(t)| \leq 1$ , for  $t \in [0, 1]$ , and  $x(0) = 1$ ,  $x(1) = \frac{1}{2}$ , where  $f(x, u) = -\frac{1}{2}x + x^2 + u(t)$ . An additional state constraint is imposed in the second example. The initial trajectory used to start the iteration was  $x^0(t) = 1$ , for  $t \in [0, 1]$ .

For these examples the simplest (forward) finite difference scheme was used, namely,

$$(5.3) \quad x_{i+1} - x_i = \Delta t f(x_i, u_i), \quad i = 0, 1, \dots, m-1,$$

so that

$$(5.4) \quad \begin{aligned} v_i = \Delta t f(x_i, u_i) + x_i - x_{i+1} &= (1 - \frac{1}{2}\Delta t)x_i + \Delta t(x_i)^2 \\ &+ \Delta t u_i - x_{i+1}, \quad i = 0, 1, \dots, m-1. \end{aligned}$$

For  $x_i^j$  known, the linearized system which must be satisfied by  $x_i^{j+1}$  and  $u_i^{j+1}$  is

$$(5.5) \quad \begin{aligned} w_i &= -x_{i+1}^{j+1} + [1 + \Delta t(2x_i^j - \frac{1}{2})]x_i^{j+1} + \Delta t u_i^{j+1} - \Delta t(x_i^j)^2 \\ &= 0, \quad i = 0, 1, \dots, m-1. \end{aligned}$$

This system is solved using the specified initial value for  $x(0)$  to give the  $x_i^{j+1}$  explicitly as linear functions of the  $u_i^{j+1}$ ,

$$(5.6) \quad x_i^{j+1} = d_i^{j+1}(u_{i-1}^{j+1}, \dots, u_0^{j+1}), \quad i = 1, \dots, m.$$

The following linear programming problem (in the dual form) is then solved at each iteration to give the new optimal control  $u_i^{j+1}$ ,

$i = 0, 1, \dots, m-1$ :

$$(5.7) \quad \min_{u_i} \left\{ \sum_{i=0}^{m-1} u_i \mid -1 \leq u_i \leq 1, i = 0, 1, \dots, m-1; \right. \\ \left. \frac{1}{2} \leq d_m^{j+1}(u_{m-1}, u_{m-2}, \dots, u_0) \leq \frac{1}{2} \right\}.$$

The corresponding state trajectory  $x_i^{j+1}$ ,  $i = 1, \dots, m$ , is then given by (5.6).

The iteration was started with  $x_i^0 = 1$ ,  $i = 0, 1, \dots, m$ , and a value of  $m = 20$  ( $\Delta t = 0.05$ ) was used. The results for the first numerical example are shown in Figs. 4 and 5. Convergence was achieved (within the desired accuracy) in three iterations. However, the difference between  $x^2$  and  $x^* = x^3$  is too small to be shown graphically (Fig. 4). Note the rapid convergence even though the initial guess,  $x_i^0$ , for the trajectory was very poor and did not even satisfy the terminal boundary condition. The corresponding optimal control  $u_i^*$  is shown in Fig. 5. The monotone behavior of the function value is verified by the successive values of  $\phi^j = \sum_{i=0}^{m-1} u_i^j$ . These were  $\phi^1 = -0.286$ ,  $\phi^2 = -0.946$ , and  $\phi^3 = -0.950$ .

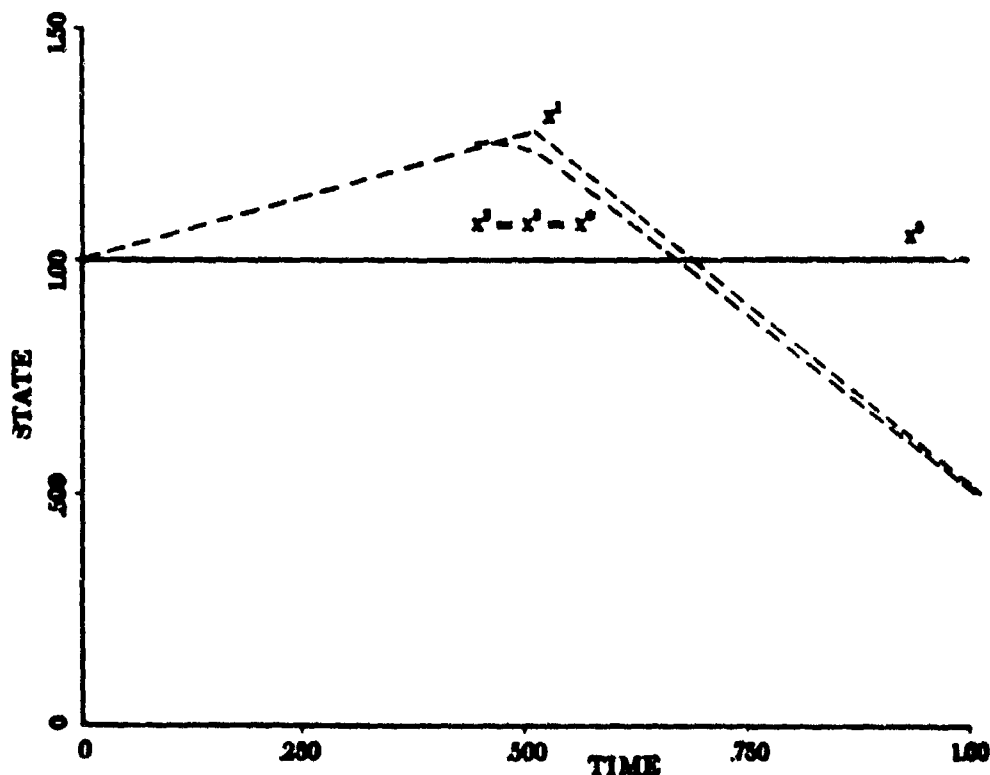


FIG. 4. Initial and optimal state trajectories for nonlinear numerical example

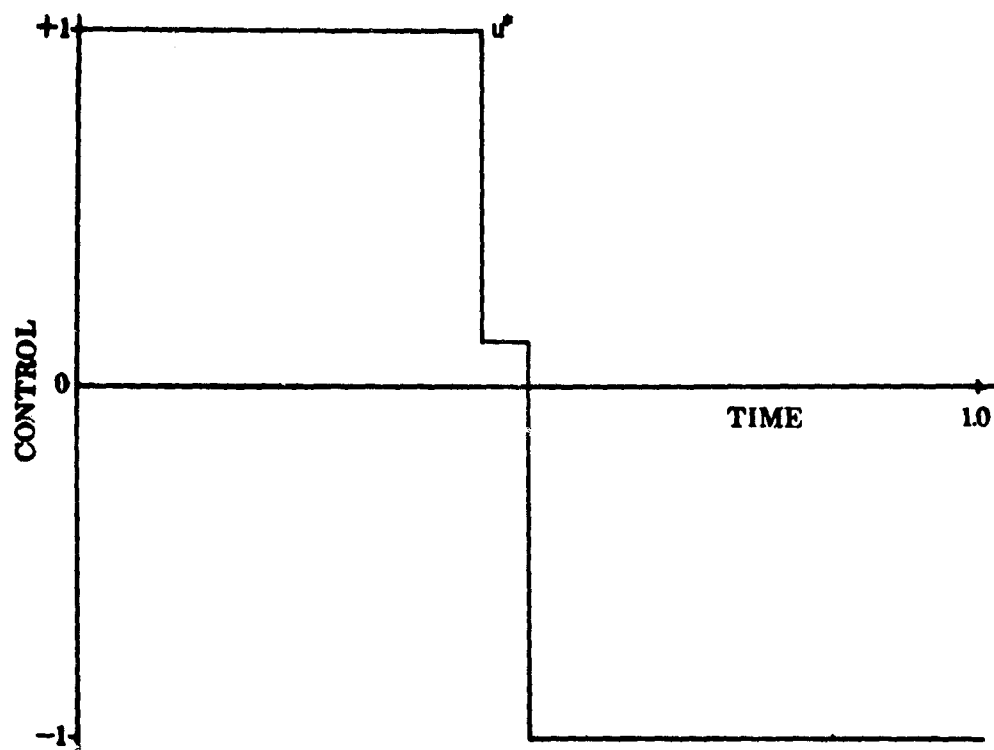


FIG. 5. Optimal control for nonlinear numerical example

For the second example the state constraint,  $x(\frac{1}{2}) \leq -\frac{1}{2}$ , was imposed. This of course eliminates the solution shown in Fig. 4. The sequence of 5 state trajectories obtained is shown in Fig. 6. The corresponding function values were  $\phi^1 = 2.792$ ,  $\phi^2 = -0.144$ ,  $\phi^3 = -0.656$ ,  $\phi^4 = -0.972$ , and  $\phi^5 = -1.008$ . The control from the first iteration  $u_i^1$  and the optimal control  $u_i^*$  are shown in Fig. 7. All of the state trajectories (except for the initial guess) are seen to satisfy the state constraints. It is interesting to observe that the method not only converges to a different trajectory  $x_i^*$  but that the added state constraint is not active for this limit trajectory. Thus the state constraint forces the solution away from its previous sequence and allows it to converge to a different local minimum. On the other hand, in some other nonlinear state-constrained cases which have been computed by this method, a state inequality constraint of the type imposed here has remained active for the limiting trajectory. Finally, it should be noted that for both cases the limiting control has the properties of an optimal control for a discrete linear problem, that is,  $n(=1)$  switchings and  $m - n(=19)$  values of  $u_i^* = \pm 1$ .

**Appendix.** In this Appendix we prove that the assumptions (1) and (2) of §3 imply the continuity of  $\Gamma y$ . We also show the validity of the modified objective function (2.13).

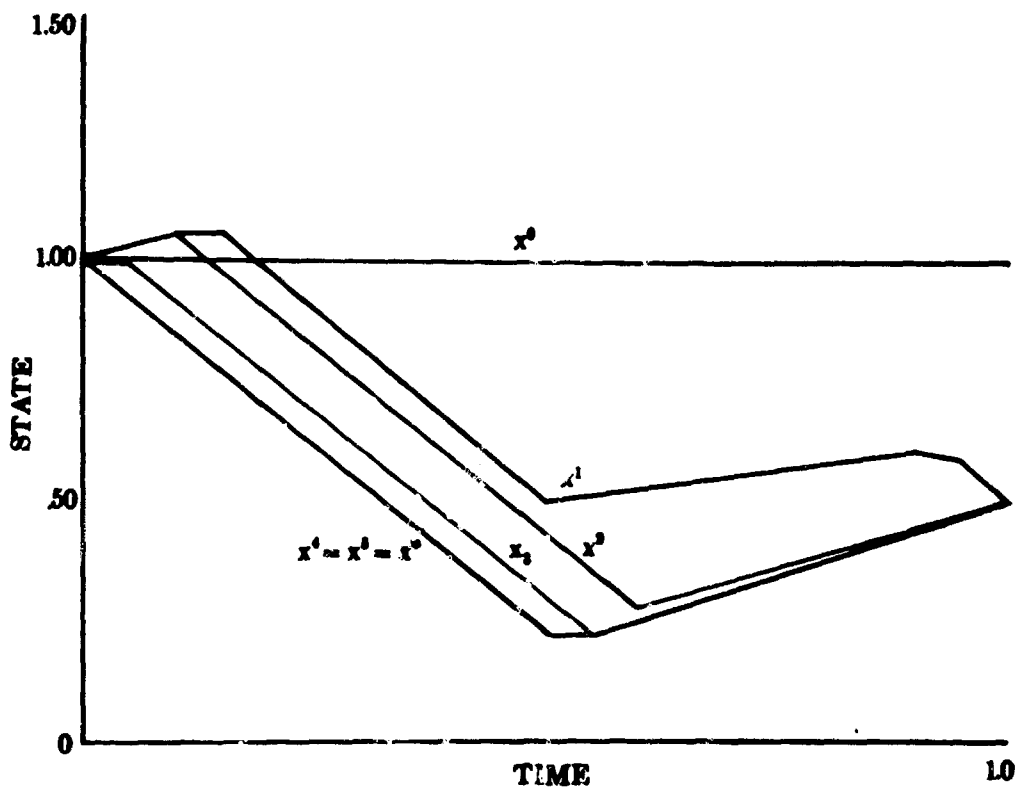


FIG. 6. State trajectories for nonlinear example with added state constraint.

We first state a condition on the rank of an augmented Jacobian matrix which insures the satisfaction of the assumption (2) of §3. In order to state this condition we must have an explicit statement of the constraints which define the compact set  $Z$ .

We will assume that  $Z$  is the polyhedral set determined by the system of  $k$  linear inequalities

$$(A.1) \quad a_i'z - b_i \geq 0, \quad i = 1, \dots, k,$$

or

$$(A.2) \quad Z = \{z \mid A'z - b \geq 0\},$$

where  $A$  is an  $s \times k$  matrix with specified columns  $a_i$ , and  $b \in E^s$  is specified. Let  $z$  denote a boundary point of  $Z$ . Then we must have at least one active constraint at  $z$ , that is,  $a_i'z - b_i = 0$  for at least one value of  $i$ . We will denote by  $\tilde{A}(z)$  the matrix whose columns represent the active constraints at  $z$ . Similarly, let  $\tilde{V}'(z)$  represent the Jacobian matrix of the vector  $\tilde{v}(z)$  which contains all components of  $v(z)$  for which  $v_i(z) = 0$ . That is,  $\tilde{v}(z) = 0$ , and  $\tilde{V}'(z) = \tilde{v}_s(z)$ .

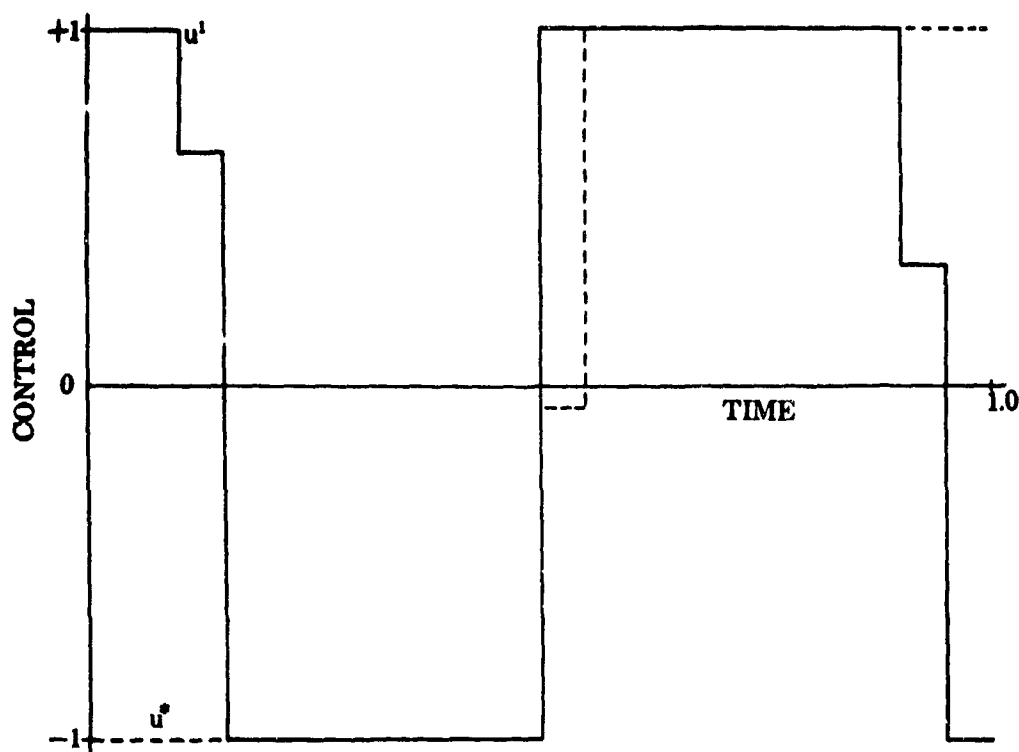


FIG. 7. Controls for nonlinear example with added state constraint

We will denote the boundary points of  $S$  by  $\partial S$ . It follows that for every  $y \in \partial S$ , the matrix

$$(A.3) \quad \tilde{B}(y) = [\tilde{V}(y) \quad \tilde{A}(y)]$$

is defined and has at least one column. We will say that  $\tilde{B}(y)$  satisfies the *full rank condition* at  $y \in \partial S$  if the columns of  $\tilde{B}(y)$  are linearly independent.

Assumption (1) implies that  $\tilde{B}(y)$  satisfies the full rank condition at every  $y \in \partial S$  which is also interior to  $Z$ . This is true because for such a point  $\tilde{A}(y) = 0$ , and  $\tilde{V}(y)$  certainly has full column rank since it consists of selected columns of  $v_s'$ . Furthermore, assumption (2) is implied by the full rank condition on  $\tilde{B}$ , as shown by the following.

**LEMMA.** *Let  $\tilde{B}(y)$  satisfy the full rank condition for every  $y \in \partial S$ . Then for each  $y \in S$ , the convex set  $\Gamma y$  contains interior points.*

*Proof.* First suppose  $\bar{y} \in S$  is an interior point of  $S$ . Since  $S \subset Z$ ,  $\bar{y}$  is an interior point of  $Z$ . Furthermore,  $w(\bar{y}, \bar{y}) = v(\bar{y}) > 0$ , so that  $\bar{y}$  is an interior point of  $W(\bar{y})$ . Therefore  $\bar{y}$  is an interior point of  $\Gamma \bar{y}$ .

Now suppose  $\bar{y} \in \partial S$ . The set  $\Gamma \bar{y}$  is the polyhedral set determined by the  $k + l$  linear inequalities

$$(A.4) \quad \Gamma \bar{y} = \{z \mid w(z, \bar{y}) \geq 0, A'z - b \geq 0\}.$$

Now consider the point  $z = \bar{y}$ . We may assume without loss of generality that

$$(A.5) \quad w_i(\bar{y}, \bar{y}) = v_i(\bar{y}) \begin{cases} = 0, & i = 1, \dots, l \leq l, \\ \geq \epsilon, & i = l+1, \dots, l, \end{cases}$$

and

$$(A.6) \quad a_i' \bar{y} - b_i \begin{cases} = 0, & i = 1, \dots, k \leq k, \\ \geq \epsilon, & i = k+1, \dots, k, \end{cases}$$

for some  $\epsilon > 0$ . Then the columns of  $\bar{V}(\bar{y})$  are the gradient vectors  $\nabla v_i(\bar{y})$ ,  $i = 1, \dots, l$ , and the columns of  $\bar{A}(\bar{y})$  are the vectors  $a_i$ ,  $i = 1, \dots, k$ . Since  $\bar{B}(\bar{y})$  satisfies the full rank condition, its columns are linearly independent and there exists no vector  $r \in E^{k+1}$ , except  $r = 0$ , such that  $\bar{B}(\bar{y})r = 0$ . Then by a variation on the Farkas lemma (see [8, Theorem 2.9, p. 48]), there exists a vector  $\bar{z} \in E^r$  such that

$$(A.7) \quad \bar{z}' \bar{B}(\bar{y}) > 0.$$

Now consider the point

$$(A.8) \quad \hat{y} = \bar{y} + \bar{z},$$

where  $\bar{z} > 0$  is chosen sufficiently small so that

$$(A.9) \quad \begin{aligned} \bar{z}' \nabla v_i(\bar{y}) &< \epsilon, & i = l+1, \dots, l, \\ \bar{z}' a_i &< \epsilon, & i = k+1, \dots, k. \end{aligned}$$

Now consider  $w_i(\hat{y}, \bar{y})$ ,  $i = 1, \dots, l$ , and  $a_i' \hat{y} - b_i$ ,  $i = 1, \dots, k$ . From (A.5), (A.6), (A.7) and (A.8) we have  $w_i(\hat{y}, \bar{y}) > 0$ ,  $i = 1, \dots, l$ , and  $a_i' \hat{y} - b_i > 0$ ,  $i = 1, \dots, k$ . From (A.5), (A.6), (A.8) and (A.9) we have  $w_i(\hat{y}, \bar{y}) > 0$ ,  $i = l+1, \dots, l$ , and  $a_i' \hat{y} - b_i > 0$ ,  $i = k+1, \dots, k$ . Therefore,  $\hat{y}$  is interior to every constraint of  $\Gamma \bar{y}$  and is an interior point of  $\Gamma \bar{y}$ .

**THEOREM 4.** *The mapping  $\Gamma y$  is continuous for  $y \in S$ .*

*Proof.* Because  $v(z) \in C^2$  on the compact set  $Z$  a uniform bound  $\gamma$  exists such that for any  $(z, y^1, y^2) \in S \times S \times S$ ,

$$(A.10) \quad \|w(z, y^1) - w(z, y^2)\| \leq \gamma \|y^1 - y^2\|.$$

Also since  $v_s(y)$  is of rank  $l$  for  $y \in S$ , the symmetric matrix  $v_s v_s'$  is positive definite at every point of  $S$ . Therefore a uniform bound  $\beta$  exists such that

$$(A.11) \quad \|(v_s v_s')^{-1}\| \leq \beta^2$$

for every  $y \in S$ .

Suppose we are given  $y^1 \in S$  and  $z^1 \in \Gamma y^1$ . Then given any  $\epsilon > 0$ , we

now show that we can choose  $\delta > 0$  so that, for each  $y^2 \in S$  with  $\|y^1 - y^2\| \leq \delta$ , we can find  $z^2 \in \Gamma y^2$  such that  $\|z^1 - z^2\| \leq \epsilon$ .

If  $z^1 \in \Gamma y^2$ , the theorem is true with  $z^2 = z^1$ . Now suppose  $z^1 \notin \Gamma y^2$ , that is, at least one component of  $w(z^1, y^2)$  is negative. Without loss of generality we assume that  $w_i(z^1, y^2) < 0$ , for  $i = 1, \dots, k \leq l$ , and  $w_i(z^1, y^2) \geq 0$ , for  $i = k+1, \dots, l$ . Since  $z^1 \in \Gamma y^1$ , we have  $w_i(z^1, y^1) \geq 0$ , for  $i = 1, \dots, l$ . Let  $\bar{w} \in E^l$  be the vector with  $\bar{w}_i = w_i(z^1, y^2) < 0$ ,  $i = 1, \dots, k$ , and  $\bar{w}_i = 0$ ,  $i = k+1, \dots, l$ . Then

$$(A.12) \quad |\bar{w}_i| \leq |w_i(z^1, y^1) - w_i(z^1, y^2)|, \quad i = 1, \dots, l,$$

so that

$$(A.13) \quad \|\bar{w}\| \leq \|w(z^1, y^1) - w(z^1, y^2)\| \leq \gamma \|y^1 - y^2\|,$$

where the last inequality follows from (A.10).

We first assume  $z^1$  is an interior point of  $Z$ . Then there is an  $\epsilon_1$  with  $0 < \epsilon_1 \leq \epsilon$ , such that  $z \in Z$  for  $\|z - z^1\| \leq \epsilon_1$ . Choose  $\delta = \epsilon_1/\beta\gamma$ , and let  $y^2$  be any point in  $S$  with  $\|y^1 - y^2\| \leq \delta$ . Now choose  $\bar{w}$  as above, and let

$$(A.14) \quad z^2 = z^1 - v_s'(y^2)[v_s(y^2)v_s'(y^2)]^{-1}\bar{w}.$$

From (3.2) we have

$$(A.15) \quad \begin{aligned} w(z^2, y^2) &= v(y^2) + v_s(y^2)[z^1 - y^2] - v_s(y^2)v_s'(y^2)[v_s(y^2)v_s'(y^2)]^{-1}\bar{w} \\ &= w(z^1, y^2) - \bar{w} \geq 0, \end{aligned}$$

so that  $z^2 \in W(y^2)$ . Furthermore from (A.14) and (A.11) we have

$$(A.16) \quad \|z^2 - z^1\|^2 = \bar{w}'[v_s(y^2)v_s'(y^2)]^{-1}\bar{w} \leq \beta^2\|\bar{w}\|^2.$$

Since  $\|y^1 - y^2\| \leq \epsilon_1/\beta\gamma$ , we get from (A.16) and (A.13) that

$$(A.17) \quad \|z^2 - z^1\| \leq \beta\|\bar{w}\| \leq \beta\gamma\|y^1 - y^2\| \leq \epsilon_1.$$

But this shows that  $z^1 \in Z$ , and therefore  $z^2 \in \Gamma y^2$ . Finally since  $\epsilon_1 \leq \epsilon$ , we have  $\|z^2 - z^1\| \leq \epsilon$ , as was to be shown.

The other possibility we must consider is that  $z^1 \in \Gamma y^1$  is a boundary point of  $Z$ . Since  $\Gamma y^1$  has interior points and is a convex set there are interior points in the neighborhood of every point of  $\Gamma y^1$  (see, for example, [7]). In particular there exist  $\epsilon_2$ ,  $0 < \epsilon_2 \leq \epsilon/2$ , and  $z^3 \in \Gamma y^1$ , such that  $\|z^3 - z^1\| \leq \epsilon/2$  and  $\|z - z^3\| \leq \epsilon_2$  implies that  $z$  is interior to  $Z$ . Now choose  $\delta = \epsilon_2/\beta\gamma$ , and replace  $z^1$  by  $z^3$  in the previous argument. This gives a point  $z^2 \in \Gamma y^2$  with  $\|z^2 - z^3\| \leq \epsilon/2$ . It follows that  $\|z^2 - z^1\| \leq \epsilon$ .

We now prove the statement about the modified objective function (2.13) made at the end of §2. We define the  $s \times (l + \bar{k})$  augmented Jacobian matrix  $B(y) = [v_s'(y) \quad \bar{A}(y)]$ . Let  $\bar{\phi}(z)$  be as in (2.13).

**THEOREM 5.** Let  $B(y)$  have full column rank for every  $y \in S$ . Then a value of  $\alpha$  exists such that every local solution of

$$(A.17) \quad \min_z \{ \bar{\phi}(z) \mid z \in Z, v(z) \geq 0 \}$$

is also a local solution of

$$(A.18) \quad \min_z \{ \phi(z) \mid z \in Z, v(z) = 0 \}.$$

*Proof.* Since  $\phi \in C^1$  and  $B(y)$  has full column rank on the compact set  $S$ , there are constants  $\alpha_1$  and  $\epsilon_1$  such that for any  $y \in S$ ,

$$(A.19) \quad \|\nabla\phi(y)\| \leq \alpha_1,$$

and

$$(A.20) \quad \|B(y)r\| \geq \epsilon_1\|r\|, \quad r \in E^{l+k}.$$

We choose  $\alpha > \alpha_1/\epsilon_1$ . Let  $y^*$  be a local minimum of (A.17). Because of the rank condition on  $B(y)$ , the necessary Kuhn-Tucker conditions are satisfied at  $y^*$ . The relevant conditions are that there exist vectors  $p \geq 0$  and  $q \geq 0$  such that

$$(A.21) \quad v_i'(y^*)p_i + \tilde{A}(y^*)q = \nabla\bar{\phi}(y^*) = \nabla\phi(y^*) + \alpha \sum_{i=1}^l \nabla v_i(y^*)$$

and

$$(A.22) \quad v_i(y^*)p_i = 0, \quad i = 1, \dots, l.$$

We let  $r' = (p_1 - \alpha, \dots, p_l - \alpha, q_1, \dots, q_k)$ , and write (A.21) as

$$(A.23) \quad B(y^*)r = \nabla\phi(y^*).$$

From (A.19) and (A.20) it follows that

$$(A.24) \quad \epsilon_1\|r\| \leq \|B(y^*)r\| = \|\nabla\phi(y^*)\| \leq \alpha_1,$$

or  $\|r\| \leq \alpha_1/\epsilon_1$ . But this requires  $|\alpha - p_i| \leq \alpha_1/\epsilon_1 < \alpha$ ,  $i = 1, \dots, l$ , or  $p_i > 0$ ,  $i = 1, \dots, l$ . Then from (A.22) we must have  $v_i(y^*) = 0$ ,  $i = 1, \dots, l$ , so that  $y^*$  is a feasible solution of (A.18).

Now suppose  $y^*$  is not a local minimum of (A.18). Then for some point  $y^1 \in Z$ , arbitrarily close to  $y^*$ , we have  $v(y^1) = 0$  and  $\phi(y^1) < \phi(y^*)$ . But then  $\bar{\phi}(y^1) < \bar{\phi}(y^*)$ , so that  $y^*$  is not a local minimum of (A.17).

#### REFERENCES

- [1] R. ARIS, *Discrete Dynamic Programming*, Blaisdell, New York, 1964
- [2] K. J. ARROW, L. HURWICZ, AND H. UZAWA, *Constraint qualifications in maximization problems*, Naval Res. Logist. Quart., 8 (1961), pp. 175-191.
- [3] R. E. BELLMAN, I. GLICKSBERG, AND O. A. GROSS, *Some aspects of the mathe-*

- mathematical theory of control processes*, R-313, The RAND Corporation, Santa Monica, 1958.
- [4] R. E. BELLMAN AND R. KALABA, *Dynamic programming, invariant imbedding and quasilinearization: comparisons and interconnections*, Computing Methods in Optimization Problems, Balakrishnan and Neustadt, eds., Academic Press, New York, 1964, pp. 135-145.
  - [5] L. D. BERKOVITZ, *On control problems with bounded state variables*, J. Math. Anal. Appl., 5 (1962), pp. 488-498.
  - [6] A. E. BRYSON, W. F. DENHAM, F. J. CARROLL, AND K. MIKAMI, *Lift or drag programs that minimize re-entry heating*, J. Aerospace Sci., 29 (1962), pp. 420-430.
  - [7] H. G. EGGESTON, *Convexity*, Cambridge University Press, Cambridge, 1958, p. 9.
  - [8] D. GALE, *The Theory of Linear Economic Models*, McGraw-Hill, New York, 1960.
  - [9] P. HENRICI, *Discrete Variable Methods in Ordinary Differential Equations*, Wiley, New York, 1962, pp. 366-374.
  - [10] R. E. KOPP AND R. MCGILL, *Several trajectory optimization techniques*, Computing Methods in Optimization Problems, Balakrishnan and Neustadt, eds., Academic Press, New York, 1964, pp. 65-89.
  - [11] H. W. KUHN AND A. W. TUCKER, *Nonlinear programming*, Proceedings of Second Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley, 1951, pp. 481-492.
  - [12] C. W. MERRIAM, *An algorithm for the iterative solution of a class of two-point boundary value problems*, this Journal, 2 (1964), pp. 1-10.
  - [13] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE, AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.
  - [14] J. B. ROSEN, *Optimal control and convex programming*, Nonlinear Programming--A Course, J. Abadie, ed., North-Holland, Amsterdam, to appear.

## Numerical Solution of Optimal Control Problems

J. B. Rosen

Computer Sciences Department  
University of Wisconsin

### Summary

I. The discrete optimal control problem to be considered is as follows:

Let  $x_1 \in E^n$  denote the state vector at time  $t_1$ , and  $u_1 \in E^r$  the corresponding control vector. The system dynamics are given by

$$x_{i+1} = x_i + f(x_i, u_i), \quad i = 0, \dots, m-1 \quad (1.1)$$

where the controls  $u_i$  must be selected so that

$$\begin{aligned} u_i &\in U_i \subset E^r, \quad i = 0, 1, \dots, m-1 \\ x_i &\in X_i \subset E^n, \quad i = 0, 1, \dots, m \end{aligned} \quad (1.2)$$

It is assumed that the sets  $U_i$  and  $X_i$  are compact and convex, and that  $f$  is continuous on  $U_i \times X_i$ . We call the sequence  $\{x_i\}$  the state trajectory, the sequence  $\{u_i\}$  the control, and we denote by  $z = \{x_i, u_i\}$  the direct product of these two sequences. We say that  $z$  is admissible if  $\{x_i\}$  and  $\{u_i\}$  satisfy (1.1) and (1.2). Note that we can specify the initial and terminal values  $x_0$  and  $x_m$  by setting  $X_0 = x_0$  and  $X_m = x_m$ .

We assume  $\sigma$  to be continuous on  $U_1 \times X_1$ , and define

$$\varphi(z) \equiv \sum_{i=0}^{m-1} \sigma(x_i, u_i) \quad (1.3)$$

We wish to find an admissible  $z^*$  such that  $\varphi(z)$  attains its minimum, over all admissible  $z$ , at  $z = z^*$ .

Now let us consider the following continuous optimal control problem. Let  $x(t)$  and  $u(t)$  satisfy

$$\left. \begin{aligned} \dot{x} &\equiv \frac{dx}{dt} = \bar{f}(x, u) \\ u(t) &\in U(t) \\ x(t) &\in X(t) \end{aligned} \right\} \quad t \in [0, T] \quad (1.4)$$

Find  $x^*(t)$  satisfying (1.4) such that

$$\varphi[u] = \int_0^T \bar{\sigma}(x(t), u(t)) dt \quad (1.5)$$

attains its minimum over all  $x(t)$  and  $u(t)$  which satisfy (1.4). Suppose that we choose a finite difference step  $\Delta t = T/m$ , and use the simplest approximation  $\dot{x}(i\Delta t) = (x_{i+1} - x_i)/\Delta t$ . We also evaluate the integral (1.5) by the trapezoidal rule and let  $f = \Delta t \bar{f}$  and  $\sigma = \Delta t \bar{\sigma}$ . We then formally obtain the equivalent discrete problem (1.1), (1.2) and (1.3).

The terminal time  $T$  is assumed to be specified in the continuous problem as given by (1.4) and (1.5). We can however put a problem with variable terminal time into this fixed time formulation by introducing an additional state variable. To illustrate this, suppose the variable time problem is given by

$$\frac{dy}{d\tau} = g(y, u), \quad y, g \in E^{n-1}$$

$$\int_0^T \eta(y, u) d\tau = \min, \quad T \text{ variable}$$

We introduce a new state variable  $\xi > 0$ , and let  $\tau = \xi t$ ,  $t \in [0, 1]$ .

We require that  $\xi$  satisfy  $\dot{\xi} = 0$ , and that its initial value  $\xi(0)$  be determined. If we define the vectors in  $E^n$ ,

$$x = \begin{pmatrix} y \\ \xi \end{pmatrix}, \quad \bar{f} = \begin{pmatrix} \xi g \\ 0 \end{pmatrix}$$

and let  $\tau(x, u) = \xi \eta(y, u)$ , the resulting problem given by (1.4) and (1.5) with  $T = 1$  is equivalent to the variable time problem.

It should also be remarked that explicit dependence on  $t$  of  $\bar{v}$  and  $\bar{f}$  can be handled with no essential difficulty. Such dependence leads to functions  $\sigma_i$  and  $f_i$  in (1.1) and (1.3) which depend explicitly on the index  $i$ . To simplify the presentation, we will not consider such dependence.

II. We will now show that the discrete optimal control problem can be considered as a mathematical programming problem (in general, nonlinear) with a special structure [1]. We let  $s = mr + (m+1)n$ , and consider the vector  $z$  in the product space  $E^s$ . We denote by  $Z \subset E^s$ , the compact, convex subset

$$Z = \left\{ z \left| \begin{array}{l} u_i \in U_i, \quad i = 0, 1, \dots, m-1 \\ x_i \in X_i, \quad i = 0, 1, \dots, m \end{array} \right. \right\} \quad (2.1)$$

We also define a vector mapping  $v: E^s \rightarrow E^{mn}$ , so that the recursion

relations (1.1) are all given by  $v(z) = 0$ . That is we define

$$v_{i+1} \equiv f(x_i, u_i) + x_i - x_{i+1}, \quad i = 0, 1, \dots, m-1 \quad (2.2)$$

and let  $v_i$  represent the components  $(i-1)n+1, \dots, in$ , of  $v$ . The discrete optimal control problem can now be stated as that of finding a  $z^*$  which solves the mathematical programming problem

$$\min_z \left\{ \varphi(z) \mid \begin{array}{l} z \in Z \\ v(z) = 0 \end{array} \right\} \quad (2.3)$$

where  $\varphi(z)$  is given by (1.3).

The admissible (feasible) set  $S \subset E^s$  is given by

$$S = \left\{ z \mid \begin{array}{l} z \in Z \\ v(z) = 0 \end{array} \right\} \quad (2.4)$$

The set  $S$  may be empty, in which case no control and corresponding trajectory exist which satisfy (1.1) and (1.2). In many practical situations the existence of an admissible control and trajectory with the given dynamics and imposed constraints is the primary question. If no admissible solution exists it is necessary to relax the control constraints (by increasing the allowable range on some of the controls, for example) or relax the state constraints (by increasing the size of the target set  $X_m$ , for example) before the determination of an optimum solution can be considered. In some cases an admissible solution may also be achieved by an appropriate modification of the system dynamics. In any event, the determination of whether or not an admissible solution exists has been reduced to finding any feasible solution to the problem (2.3).

If  $S$  is not empty, it is a compact set, so that  $\varphi$  attains its minimum on  $S$ . If the null space of  $v$  is convex, then  $S$  is also convex.

If  $\varphi$  is convex on  $Z$  ( $\sigma$  convex on  $X_1 \times U_1$ ), then (2.3) is a convex programming problem for which both necessary and sufficient optimality conditions can be stated, and for which efficient computational methods of solution are available. It follows from (2.2) that for linear  $f$ ,  $f = Ax + Bu + q$ , where  $A$  and  $B$  are matrices and  $q$  is a constant vector, the null space of  $v$  will be a linear manifold (and therefore convex). Thus linear dynamics and a convex functional lead to a reasonably well-understood convex problem. In general, however, if  $f$  is not linear, the set  $S$  will not be convex. Necessary optimality conditions will be given for  $S$  nonconvex, but in general, conditions which are also sufficient are not known for such problems. Furthermore, for nonconvex  $S$  a problem may have many constrained local minima even with  $\varphi$  linear. Thus even if a method finds a local minimum such a minimum may be far from the desired global minimum.

III. We will now consider optimality conditions for the problem (2.3), and use these to obtain the adjoint equations and a "minimum principle" for the discrete optimal control problem. We assume that  $\varphi$  and  $v$  are in  $C^1(Z)$ , and that  $S$  is nonempty. We denote by  $v_z$  the Jacobian matrix of a function  $v$ , and  $p$  transpose by  $p'$  so that  $p'v$  denotes an inner product.

#### Sufficiency Theorem

Let  $\varphi$  be convex and  $v$  be linear on  $Z$ . A sufficient condition that  $z^* \in S$  solves (2.3) is that there exists  $p \in E^{mn}$  such that

$$[\varphi_z(z^*) + p'v_z(z^*)](z-z^*) \geq 0, \quad \forall z \in S \quad (3.1)$$

Proof: Since  $\varphi$  is convex and  $v$  linear, the function  $\psi = \varphi + p'v$  is convex on  $Z$ . Then for every  $z \in Z$ ,  $\psi(z) - \psi(z^*) \geq \psi_z(z^*)[z-z^*] \geq 0$ , by (3.1). Since  $z^* \in S$ ,  $v(z^*) = 0$ , so that we have

$$\varphi(z) - \varphi(z^*) \geq -p'v(z) = 0, \quad \forall z \in S.$$

Thus,  $z^*$  is a global minimum on  $S$ .

### Necessity Theorem

Assume that the compact, convex set  $Z$  has interior points. Let  $z^*$  solve (2.3). Then there exists a scalar  $\mu \geq 0$ , and  $p \in E^{mn}$ , not both zero, such that

$$\Phi_z(z^*)(z-z^*) \geq 0, \quad \forall z \in S \quad (3.2)$$

where

$$\Phi(z) \equiv \mu \varphi(z) + p'v(z) \quad (3.3)$$

The proof of this theorem is too long to be included here, and is given in [2]. It should also be noted that if an appropriate constraint qualification is satisfied we can choose  $\mu = 1$ .

If we restrict the functions  $\varphi$  and  $v$  as in the sufficiency theorem, we obtain a Minimum Principle.

Let  $\varphi$  be convex and  $v$  linear on  $Z$ , and let  $z^*$  solve (2.3). Then there exist multipliers  $\mu \geq 0$ , and  $p$ , not both zero, such that  $\Phi(z)$  attains its minimum over  $z \in S$  at  $z^*$ , where  $\Phi(z)$  is given by (3.3). The proof follows immediately from the convexity of  $\Phi$  and (3.2).

We are now in a position to apply these results directly to the discrete optimal control problem. If we denote by  $p_i \in E^n$ , the multipliers corresponding to  $v_i$ , we obtain from (1.3) and (3.3).

$$\Phi(z) = \mu \sum_{i=0}^{m-1} \sigma(x_i, u_i) + \sum_{i=0}^{m-1} p'_{i+1} [f(x_i, u_i) + x_i - x_{i+1}] \quad (3.4)$$

We will let  $f^*_{x1} \equiv f_{x1}(x^*_1, u^*_1)$ , etc., and  $p_0 \equiv 0$ . Then the necessary condition (3.2) can be written

$$[\mu \sigma^*_{x1} + p'_{i+1} (I + f^*_{x1}) - p'_i] (x_i - x^*_i) \geq 0, \quad \forall x_i \in X_i, \quad i=0, 1, \dots, m-1 \quad (3.5)$$

$$-p'_m (x_m - x^*_m) \geq 0, \quad \forall x_m \in X_m \quad (3.6)$$

$$[\mu \sigma^*_{u1} + p'_{i+1} f^*_{u1}] (u_i - u^*_i) \geq 0, \quad \forall u_i \in U_i, \quad i=0, 1, \dots, m-1 \quad (3.7)$$

If we assume that the initial state is specified as a point, i.e.,  $X_0 = \bar{x}_0$ , then  $x^*_0 = \bar{x}_0$  and (3.5) is satisfied for  $i = 0$ . Now suppose further that no state constraints are active except for the terminal constraint, i.e.,  $x^*_i \in \text{int } X_i$ ,  $i = 1, \dots, m-1$ . Then (3.5) requires that the expression  $[\cdot]$  vanish. That is

$$p'_i - p'_{i+1} = p'_{i+1} f^*_{x1} + \mu \sigma^*_{x1}, \quad i = m-1, m-2, \dots, 0. \quad (3.8)$$

The multipliers  $p_i$  are therefore determined by the recursion relation (3.8) starting with a vector  $p_m$  satisfying (3.6). For  $x^*_m$  on the boundary of  $X_m$ ,  $p_m$  must be parallel to the (outward) normal vector to a supporting hyperplane at  $x^*_m$ . The recursion relation (3.8) is seen to be a finite difference approximation to the adjoint differential equation

$$-\dot{p}' = p' \bar{f}_x + \mu \bar{\sigma}_x \quad (3.9)$$

for the continuous optimal control problem.

Let us now define

$$H(x, u, p, q) = p'f(x, u) + (p-q)'x + \mu \sigma(x, u) \quad (3.10)$$

By rearranging terms we obtain from (3.4)

$$\Phi(z) = \sum_{i=0}^{m-1} H(x_i, u_i, p_{i+1}, p_i) - p_m'x_m \quad (3.11)$$

For  $\sigma$  convex and  $f$  linear on  $X_i \times U_i$ , the previous minimum principle applies. The optimal trajectory  $\{x_i^*\}$  and control  $\{u_i^*\}$  must therefore satisfy the following

#### Discrete Minimum Principle

If  $\{x_i^*\}$  and  $\{u_i^*\}$  are optimal, then

$$H(x_i^*, u_i^*, p_{i+1}, p_i) \leq H(x_i, u_i, p_{i+1}, p_i), \quad \forall x_i \in X_i, u_i \in U_i \quad (3.12)$$

$i=0, 1, \dots, m-1$

where the adjoint vectors  $p_i$  are determined by (3.6) and (3.8).

IV. In order to describe the computational solution of the discrete optimal control problem we first consider a linear recursion relation with

$f = Ax + Bu$ , and also give a more explicit statement of the constraint

sets  $X_i$  and  $U_i$ . To simplify the discussion we will assume that  $X_0 = \bar{x}_0$ , and that the  $X_i$  and  $U_i$  are polyhedral sets defined by the linear inequalities

$$\begin{aligned} X_i &= \{ x \mid G_i'x - \bar{g}_i \leq 0 \}, \quad i = 1, \dots, m \\ U_i &= \{ u \mid H_i'u - \bar{h} \geq 0 \}, \quad i = 0, \dots, m-1 \end{aligned} \quad (4.1)$$

where  $G_i$  and  $H$  are constant matrices and  $\bar{g}_i$  and  $\bar{h}$  are constant vectors. If these sets are just upper and lower bounds we have

$$G_1 = (I_n \ -I_n) \text{ and } H = (I_r \ -I_r).$$

The problem (1.1) - (1.3) now becomes

$$\min_{x_1, u_1} \left\{ \begin{array}{l} \sum_{i=0}^{m-1} \sigma(x_i, u_i) \\ x_0 = \bar{x}_0 \\ x_{i+1} = (I+A)x_i + Bu_i, \quad i = 0, \dots, m-1 \\ G'_1 x_1 - \bar{q}_1 \geq 0, \quad i = 1, \dots, m \\ H'u_1 - \bar{h} \geq 0, \quad i = 0, \dots, m-1 \end{array} \right\} \quad (4.2)$$

If  $\sigma$  is convex, this is a convex programming problem with linear equality and inequality constraints. It consists of  $m(n+r)$  variables,  $mn$  equality constraints, and  $2m(n+r)$  inequality constraints if they are bounds. This could be solved directly, but will be a large problem if  $m$  is large. A more efficient method of solution is to use the linear equations to eliminate the state vectors  $x_i$  and map the entire problem into the control space.

This is done by means of the following

#### Lemma

Let  $x_i$  satisfy the recursion relation

$$x_{i+1} = K_1 x_i + b_i, \quad i = 0, 1, \dots, m-1 \quad (4.3)$$

with  $x_0$  specified and  $K_1$  nonsingular. Then

$$x_i = Y_i x_0 + Y_i \sum_{j=0}^{i-1} \Lambda'_{j+1} b_j, \quad i = 1, \dots, m \quad (4.4)$$

where the matrices  $Y_i$  and  $\Lambda_i$  satisfy

$$Y_{i+1} = K_1 Y_i, \quad Y_0 = I, \quad i = 0, 1, \dots, m-1 \quad (4.5)$$

and

$$\Lambda'_i = \Lambda'_{i+1} K_1, \quad \Lambda'_m = Y_m^{-1}, \quad i = m-1, \dots, 1 \quad (4.6)$$

If we apply this to the equality constraints with  $K_1 = I + A$ , and

$b_1 = Bu_1$  we obtain

$$x_1 = Y_1 x_0 + \sum_{j=0}^{i-1} Q'_{1j} u_j, \quad i = 1, \dots, m \quad (4.7)$$

where  $Q'_{1j} = Y_1 \Lambda'_{j+1} B_1$ . The important point to note is that given the matrices  $A$  and  $B$ , we can explicitly compute the matrices  $Y_1$  and  $Q'_{1j}$  by operations with  $(n \times n)$  matrices. Since  $n$  (the dimensionality of the state vector) is usually small compared to  $m$ , this is an efficient computation.

We now use (4.7) to eliminate the  $x_1$  from the state constraints and  $\sigma$  in (4.2). The state constraints can then be written

$$\sum_{j=0}^{i-1} D'_{1j} u_j - d_1 \geq 0, \quad i = 1, \dots, m \quad (4.8)$$

where  $D'_{1j} = G'_1 Q'_{1j}$  and  $d_1 = \bar{g}_1 - G'_1 Y_1 x_0$ . A similar substitution in  $\sigma$  gives an objective function  $\rho(u)$  depending on the controls  $u = \{u_1\}$ .

only. If  $\sigma$  is convex,  $\rho(u)$  will also be convex because of the linearity of (4.7). The problem (4.2) has therefore been reduced to

$$\min_{u_1} \left\{ \begin{array}{l} \rho(u) \\ \sum_{j=0}^{i-1} D'_{1j} u_j \geq d_1, \quad i = 1, \dots, m \\ H^0 u_1 \geq \bar{h}, \quad i = 0, \dots, m-1 \end{array} \right\} \quad (4.9)$$

This reduced problem has  $mr$  variables, no equalities, and the same number of inequalities as (4.2). In most control problems  $r < n$ , so that we have cut the problem size at least in half. Since there are more inequality constraints than variables, and no nonnegativity requirements, this is best

solved by a convex method in the dual space (such as gradient projection). Computational efficiency is also improved by taking advantage of the upper triangular structure of the first  $2mn$  constraints and the block diagonal structure of the remaining  $2mr$  constraints.

In the important case where  $\sigma$  is linear, (4.9) should be considered as the unsymmetric dual problem. The equivalent primal, with  $mr$  rows is then efficiently solved by any standard LP routine, which of course gives the desired dual variables (the controls) as the elements of the pricing vector. For a more complete discussion of the linear recursion relation problem, see [1].

V. The method of the previous section can only be applied directly when the system dynamics are described by a linear recursion relation. However, with appropriate convexity requirements on  $f$ , the nonlinear problem can be solved by an iterative solution of linearized problems. At each iteration the function  $f$  is linearized about the previous state and control, and a linearized problem of the kind discussed above is solved. This method is fully described in [3].

REFERENCES

1. J. B. Rosen, "Optimal control and convex programming", Proc. IBM Symp. on Control Theory and Appl., Yorktown Heights, N.Y. (Oct. 1964) pp. 223-237.
2. O. L. Mangasarian, Nonlinear Programming, McGraw Hill (to be published in 1968), Chap. 11.
3. J. B. Rosen, "Iterative solution of nonlinear optimal control problems", J. SIAM Control 4 (1966), pp. 223-244.

# COMPLEMENTARY SLACKNESS IN DUAL LINEAR SYSTEMS

The first part of this paper develops a convenient technique for working with a pair of dual (i.e., complementary orthogonal) linear subspaces,  $\Xi$  and  $X$ , in a linear space of  $n$ -tuples (from an ordered field). The second part deals with the fundamental existence theorem that the dual subspaces  $\Xi$  and  $X$  must contain  $n$ -tuples  $\xi$  and  $x$  such that  $\xi_i \geq 0$ ,  $x_i \geq 0$ ,  $\xi_i x_i = 0$ , and  $\xi_i + x_i > 0$  for  $i = 1, \dots, n$ . Applications of this "complementarity slackness" are given.

$$\begin{array}{c}
 \begin{array}{c} (x) \\ x_1 \quad \dots \quad x_n \\ \lambda_1 \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & (A) & \vdots \\ a_{m1} & \dots & a_{mn} \end{bmatrix} \begin{array}{l} =0 \\ \vdots \\ =0 \end{array} \\ \vdots \\ \lambda_m \\ \hline =\xi_1 \quad \dots \quad =\xi_n \\ (= \xi) \end{array} \\
 \end{array}
 \qquad
 \begin{array}{c}
 \begin{array}{c} (\bar{x}) \\ x_{\bar{r}+1} \quad \dots \quad x_{\bar{n}} \\ \xi_{\bar{1}} \begin{bmatrix} m_{11} & \dots & m_{1s} \\ \vdots & (M) & \vdots \\ m_{r1} & \dots & m_{rs} \end{bmatrix} \begin{array}{l} =-x_{\bar{1}} \\ \vdots \\ =-x_{\bar{r}} \end{array} \\ \vdots \\ \xi_{\bar{r}} \\ \hline =\xi_{\bar{r}+1} \quad \dots \quad =\xi_{\bar{n}} \\ (= \bar{\xi}) \end{array} \\
 \end{array}
 \end{array}
 \quad (= -\bar{\xi})$$

( $m=2, n=4; r=2, s=2$ ) example ( $\bar{1}=2, \bar{2}=3$  and  $\bar{3}=1, \bar{4}=4$ )

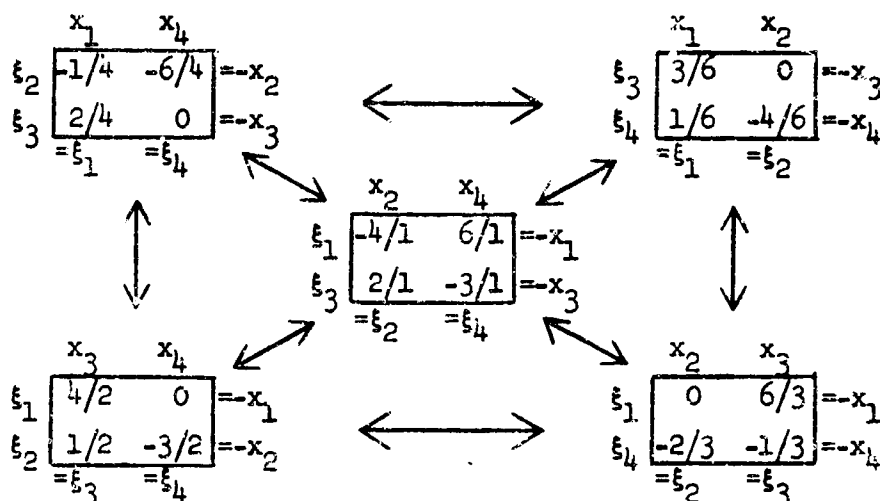
$$\begin{array}{c}
 \begin{array}{c} x_1 \quad x_2 \quad x_3 \quad x_4 \\ \lambda_1 \begin{bmatrix} 1 & 2 & 3 & -3 \\ -1 & 6 & 1 & -9 \end{bmatrix} \begin{array}{l} =0 \\ =0 \end{array} \\ \vdots \\ \lambda_2 \\ \hline =\xi_1 \quad =\xi_2 \quad =\xi_3 \quad =\xi_4 \\ \text{(standard form)} \end{array} \\
 \end{array}
 \qquad
 \begin{array}{c}
 \begin{array}{c} x_1 \quad x_4 \\ \xi_2 \begin{bmatrix} -1/4 & -6/4 \\ 2/4 & 0 \end{bmatrix} \begin{array}{l} =-x_2 \\ =-x_3 \end{array} \\ \vdots \\ \xi_3 \\ \hline =\xi_1 \quad =\xi_4 \\ \text{(canonical form)} \end{array} \\
 \end{array}$$

The above  $\Lambda$ -schema exhibits two dual systems of homogeneous linear equations in (Dantzig's) "standard form." The "greek" system  $\Lambda A = \xi$  determines the row-space of the matrix  $A$ , i.e., the linear subspace  $\Xi$  of all row-vectors  $\xi$  expressible, via parameters  $\lambda$ , as linear combinations of the rows of  $A$ . The "latin" system  $Ax = 0$  determines the orthogonal complement in  $n$ -space of the row-space of  $A$ , i.e., the linear subspace  $X$  of all column-vectors  $x$  orthogonal to each row of  $A$  — and therefore to each  $\xi$  of  $\Xi$ , since  $\xi x = (\Lambda A)x = \lambda(Ax) = 0$ . Let rank  $A = r$ ; then  $\dim \Xi = r$  and  $\dim X = n - r = s$ . (Any matrix  $\bar{A}$  that is row-equivalent to  $A$  (including insertion or deletion of rows of zeros) yields an  $\bar{\Lambda}$ -schema, with its own parametric  $\bar{\lambda}$ , which determines the same dual subspaces  $\Xi$  and  $X$ .)

Now partition  $Ax = 0$  into  $\bar{A}\bar{x} + \bar{A}'\bar{x}' = 0$ , where the submatrix  $\bar{A}$  consists of  $r$  linearly independent columns of  $A$ . (There are at most  $n!/r!$  such partitions.) Since  $\bar{A}$  provides a basis for the columns of  $A$ , there exists an  $r$  by  $s$  matrix  $M$  such that the submatrix  $\bar{A}' = \bar{A}M$ . Gauss-Jordan (complete) elimination reduces  $\bar{A}\bar{x} + \bar{A}'\bar{x}' = 0$  to  $\bar{I}\bar{x} + M'\bar{x}' = 0$ , i.e.,  $M'\bar{x}' = -\bar{x}$ , and reduces  $\Lambda A = \xi$  to  $\bar{\Lambda}\bar{x} = \bar{\xi}$ . The above  $M$ -schema exhibits these reduced systems, which are dual linear systems in (Dantzig's) "canonical form"; note that  $\bar{1}, \dots, \bar{r}$  index the (basic) columns of  $\bar{A}$  and that  $\bar{r}+1, \dots, \bar{n}$  index the (non-basic) columns of  $\bar{A}$ .

Pivot on any nonzero entry  $m_{pq}$  to exchange  $x_{\bar{p}}$  with  $x_{\bar{r}+\bar{q}}$  and  $\xi_{\bar{p}}$  with  $\xi_{\bar{r}+\bar{q}}$  (but with no other marginal changes) to get an  $\bar{M}$ -schema with entries as follows:  $\bar{m}_{pq} = 1/m_{pq}$ ,  $\bar{m}_{pj} = m_{pj}/m_{pq}$ ,  $\bar{m}_{iq} = -m_{iq}/m_{pq}$ ,  $\bar{m}_{ij} = m_{ij} - m_{iq}m_{pj}/m_{pq}$  (for each  $i \neq p$  and each  $j \neq q$ ). Note the sign change from  $m_{iq}$  to  $\bar{m}_{iq}$  if  $m_{pq} > 0$ , and from  $m_{pj}$  to  $\bar{m}_{pj}$  if  $m_{pq} < 0$  (as exemplified below). By a finite

succession of such pivot steps (and rearrangements of rows and/or columns) we can pass from any M-schema for  $\Xi$  and  $X$  to any other (in the "combinatorial equivalence" class). For example,

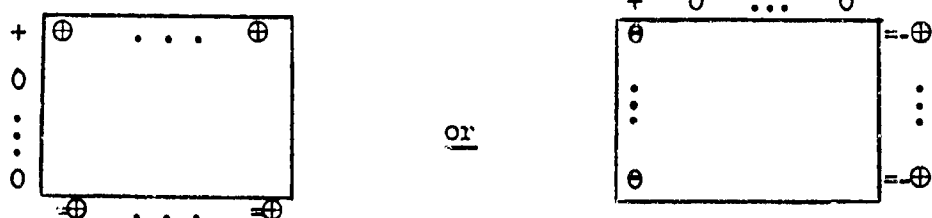


The double arrows indicate pivot steps (in either direction) and appropriate rearrangement. There are just five M-schemata in this case (not  $4!/2! = 6$ ) because the 2nd and 4th columns of the initial matrix  $A$  are linearly dependent.

\* \* \*

We now turn our attention to the (polyhedral) cones  $\Xi^*$  and  $X^*$  in which the dual subspaces  $\Xi$  and  $X$  intersect the (closed) orthant of all nonnegative  $n$ -tuples. Borrowing a convenient term from Linear Programming, we say that any  $\xi$  in  $\Xi^*$  or any  $x$  in  $X^*$  is feasible; i.e.,  $\xi$  is feasible if  $\lambda A = \xi \geq 0$  for some  $\lambda$ , and  $x$  is feasible if  $Ax = 0$  and  $x \geq 0$ . Note that  $\xi x = 0$  implies, for any feasible  $\xi$  and any feasible  $x$ , that  $\xi_i x_i = 0$ , i.e.,  $\xi_i = 0$  or  $x_i = 0$  (or both), for  $i = 1, \dots, n$ .

.. We say that a feasible  $\xi$  or  $x$  is basic if, for any partition of  $A$  into  $\bar{A}$ ,  $A$  (as specified above), exactly one component of  $\xi$  or  $x$  is positive and if  $\xi$  or  $x$  is normalized so that the sum of the components is one. Hence a basic feasible  $\xi$  or  $x$  corresponds to a nonnegative row or nonpositive column of some M-schema:

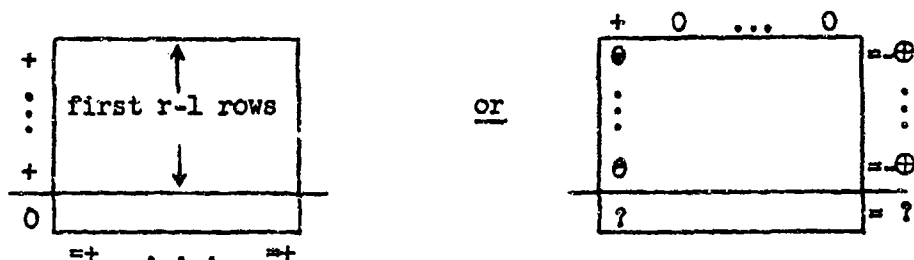


where  $+$ ,  $\oplus$ ,  $0$ ,  $\ominus$ ,  $-$  denote numbers that are positive, nonnegative, zero, non-positive, negative, respectively. In the above example  $\xi = (1/3, 0, 2/3, 0)$  and  $x^T = (0, 3/5, 0, 2/5)$  are basic feasible 4-tuples. Clearly the set of basic feasible  $\xi$ 's (or  $x$ 's) is finite, since the class of M-schemata is finite. [The reason for the term "basic" in this context is that it can be shown that the basic feasible  $\xi$ 's or  $x$ 's determine the "extreme rays" of the cone  $\Xi^*$  or  $X^*$ ; i.e., that any feasible  $\xi$  or  $x$  is expressible as a nonnegative linear combination of the basic feasible  $\xi$ 's or  $x$ 's.]

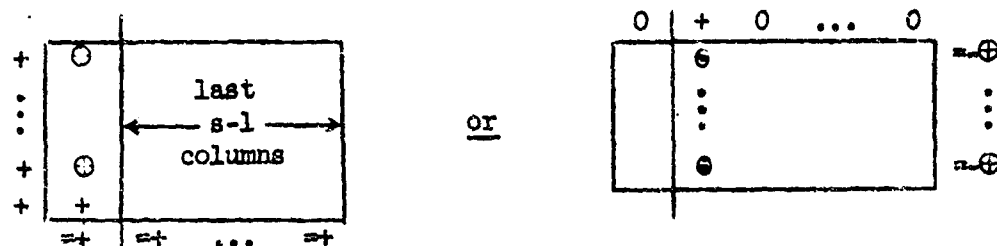
**LEMMA.** Either there is an all-positive (feasible)  $\xi$  or there is a basic feasible  $x$ .

**Proof.** Let  $L(r,s)$  assert this Lemma when  $\dim \Xi = r$  and  $\dim X = s$ . We use induction on  $r+s = n$ .  $L(1,s)$  and  $L(r,1)$  are trivial to prove.

We prove  $L(r,s)$  for  $r > 1$  and  $s > 1$ , assuming  $L(r-1, s)$  and  $L(r, s-1)$ . Using  $L(r-1, s)$  for the first  $r-1$  rows, we get



In the left alternative, perturb 0 to a small  $\epsilon > 0$  to get an all-positive  $\xi$ . In the right alternative, there is a basic feasible  $x$  showing, unless the corner entry marked ? is positive. If this corner entry is positive, pivot on it and then use  $L(r, s-1)$  on the last  $s-1$  columns to get



In the left alternative, there is an all-positive  $\xi$ , and in the right alternative, a basic feasible  $x$ .  $L(r,s)$  is now inductively established for all  $r$  and  $s$ .

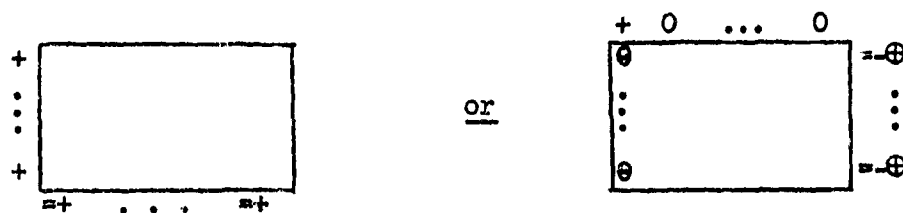
The above Lemma is essentially "the theorem of the alternative for matrices" used by J. von Neumann and O. Morgenstern to prove the Main ("Minimax") Theorem in their *THEORY OF GAMES AND ECONOMIC BEHAVIOR*. There is a dual Lemma with  $\xi$  and  $x$  interchanged (i.e., using  $-M^T$  in place of  $M$ ). The two Lemmas combine to show that there is either a basic feasible  $\xi$  or a basic feasible  $x$  (or both).

**COMPLEMENTARY SLACKNESS THEOREM.** There exists a feasible  $\xi$  and a feasible  $x$  such that  $\xi + x^T > 0$  (in all components).

In the example above,  $\xi = (1, 0, 3, 0)$  and  $x^T = (0, 3, 0, 2)$  yield  $\xi + x^T = (1, 3, 3, 2) > 0$ .

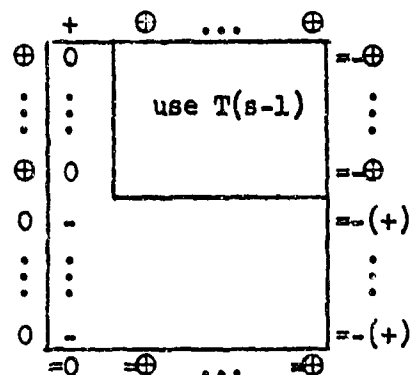
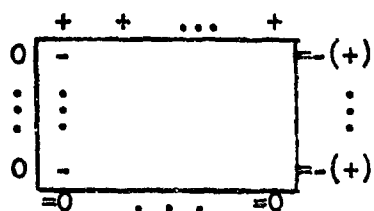
**Proof** (by T. D. Parsons). Let  $T(s)$  assert this Theorem when  $\dim X = s$ . We use induction on  $s$ .  $T(1)$  is trivial to prove.

We prove  $T(s)$  for  $s > 1$ , assuming  $T(s-1)$ . Apply  $L(r,s)$  to get



In the left alternative, taking  $x = 0$  we have  $\xi + x^T = \xi > 0$ . In the right alternative, we take the two cases

4.



In the left case, we get  $\xi = 0$  and  $x > 0$  by taking the first  $+$  at the top large enough, after the remaining  $+$ 's at the top are taken arbitrarily. In the right case, we use  $T(s-1)$  opposite the  $0$ 's in the first column and then make the first  $+$  at the top large enough. Thus, as indicated, we get  $\xi + x^T > 0$  because of  $T(s-1)$  for the components  $\oplus$ .  $T(s)$  is now inductively established for all  $s$ .

This Theorem is Theorem 1 (and 3) in Paper 1 by A. W. Tucker in LINEAR INEQUALITIES & RELATED SYSTEMS, ed. by H. W. Kuhn & A. W. Tucker (Princeton, 1956) and, with a geometric proof, "Key Theorem" in R. A. Good, "Systems of Linear Relations," SIAM REVIEW 1 (1959) 1-31. Complementary slackness refers to the existence of feasible  $\xi$  and  $x$  with "slack" ( $> 0$ ), for each column of  $A$ , in either  $\xi_j$  or  $x_j$  (but not both). The following are Corollaries (see papers cited above for references):

1. (Gordan, 1873)  $Ax = 0$  has a solution  $x \geq 0$  and  $\neq 0$  iff  $\lambda A > 0$  for no  $\lambda$ .
2. (Stiemke, 1915)  $Ax = 0$  has an all-positive solution  $x$  iff  $\lambda A \geq 0$  and  $\neq 0$  for no  $\lambda$ .
3. (Farkas, 1902)  $b \geq 0$  for all  $\lambda$  such that  $\lambda A \geq 0$  iff  $Ax = b$  for some  $x \geq 0$ .
4.  $cx \geq 0$  for all  $x \geq 0$  such that  $Ax = 0$  iff  $\lambda A + c \geq 0$  for some  $\lambda$ .
5. If  $M$  is skew-symmetric (i.e.,  $M^T = -M$ ), then  $\xi M = \eta^T$  has a solution  $\xi \geq 0$  such that  $\xi + \eta > 0$ .

Corollary 2 (= Theorem 5 in Paper 1 cited above) can be used to establish the duality and existence theorems of Linear Programming (see A. J. Goldman and A. W. Tucker, "Theory of Linear Programming," LINEAR INEQUALITIES & RELATED SYSTEMS, pp. 53-62, and R. A. Good's paper cited above, pp. 17-21).

# FINDING THE POINT OF A CONVEX POLYHEDRON NEAREST A GIVEN POINT

Problem: To minimize half the square of the distance from the point  $(2,3,-1)$  to the (solid) tetrahedron determined by

$$w = -x - y - z + 3 \geq 0, \quad x \geq 0, \quad y \geq 0, \quad z \geq 0.$$

The "objective function" to be minimized subject to these constraints is

$$f(x,y,z) = \frac{1}{2}[(x-2)^2 + (y-3)^2 + (z+1)^2] = 7 - 2x - 3y + z + \frac{1}{2}(x^2 + y^2 + z^2).$$

A necessary and sufficient condition that a point  $(x,y,z)$  of the tetrahedron yield the required minimum is that the gradient of  $f$  at the point  $(x,y,z)$  be expressible as a nonnegative linear combination of the inward normals to the constraints that are "active" at the point  $(x,y,z)$ . That is,

$$\nabla f = \begin{bmatrix} x-2 \\ y-3 \\ z+1 \end{bmatrix} = \omega \begin{bmatrix} -1 \\ -1 \\ -1 \end{bmatrix} + \xi \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + \eta \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + \zeta \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

where  $\omega, \xi, \eta, \zeta$  are each nonnegative and must be zero if the corresponding constraint is "slack" ( $> 0$ ) at the point  $(x,y,z)$ . Hence

$$\omega \geq 0, \quad \xi = \omega + x - 2 \geq 0, \quad \eta = \omega + y - 3 \geq 0, \quad \zeta = \omega + z + 1 \geq 0$$

and

$$\omega w + \xi x + \eta y + \zeta z = 0.$$

Another way of getting the same information is to form the Lagrangian function

$$\phi = f - \omega w - \xi x - \eta y - \zeta z$$

with nonnegative Lagrange multipliers  $\omega, \xi, \eta, \zeta$  and take  $\nabla \phi = 0$  with  $\phi = f$  (i.e.,  $\omega w + \xi x + \eta y + \zeta z = 0$ ). Substituting for  $f, w, \xi, \eta, \zeta$  from above, we get

$$\begin{aligned} \phi(x,y,z) &= 7 - 2x - 3y + z + \frac{1}{2}[x^2 + y^2 + z^2] - \omega(3-x-y-z) - (\omega+x-2)x - (\omega+y-3)y - (\omega+z+1)z \\ &= 7 - 3\omega - \frac{1}{2}[x^2 + y^2 + z^2]. \end{aligned}$$

Hence

$$f + \phi = -3\omega - 2x - 3y + z + 14.$$

The above linear equations for  $w, \xi, \eta, \zeta$ , and  $f + \phi$  are exhibited most conveniently in the following schema:

$\omega$	0	1	1	1	-3
$x$	-1	1*	0	0	-2
$y$	-1	0	1	0	-3
$z$	-1	0	0	1	1
1	3	-2	-3	1	14
	$=w$	$=\xi$	$=\eta$	$=\zeta$	$=f+\phi$

In addition  $w, x, y, z$  and  $\omega, \xi, \eta, \zeta$  are all to be nonnegative and

$$f - \phi = \omega w + \xi x + \eta y + \zeta z = 0.$$

We seek, therefore, a solution of the linear system in the above schema that is nonnegative in all eight variables (excluding  $f+\phi$ ) and is such that  $\omega = 0$  or/and  $w = 0, \xi = 0$  or/and  $x = 0, \eta = 0$  or/and  $y = 0, \zeta = 0$  or/and  $z = 0$ . For this purpose we turn to schemata, such as the following, that are equivalent to the above schema under principal pivoting:

2.

$\omega$	1	-1	1	1	-1
$\xi$	-1	1	0	0	-2
$y$	-1	0	1*	0	-3
$z$	-1	0	0	1	1
1	1	2	-3	1	10
	$=w$	$=x$	$=y$	$=z$	$=f+\phi$

$\omega$	3	-1	-1	-1	1
$\xi$	-1	1	0	0	-2
$\eta$	-1	0	1	0	-3
$\zeta$	-1	0	0	1	1
1	-1	2	3	-1	0
	$=w$	$=x$	$=y$	$=z$	$=f+\phi$

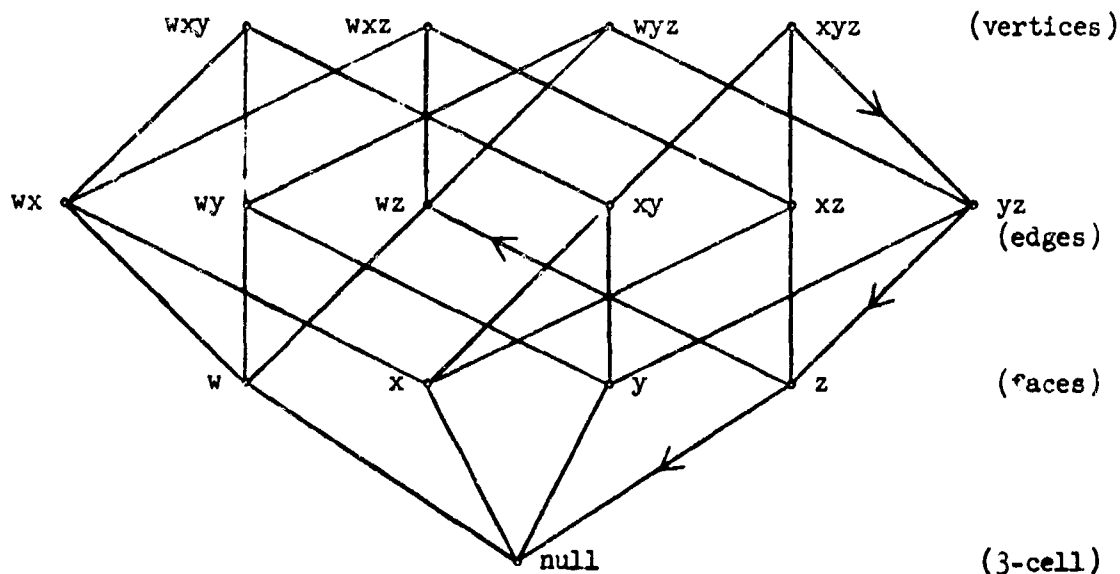
$\omega$	2	-1	-1	1	2
$\xi$	-1	1	0	0	-2
$\eta$	-1	0	1	0	-3
$z$	-1	0	0	1*	1
1	-2	2	3	1	1
	$=w$	$=x$	$=y$	$=z$	$=f+\phi$

$\xi$	1/2	-1/2	1/2	1/2	-1
$\eta$	-1/2	1/2	1/2	1/2	-2
$w$	-1/2	-1/2	1/2	1/2	1
$z$	-1/2	-1/2	1/2	3/2	2
1	1	2	1	2	3
	$=x$	$=y$	$=w$	$=z$	$=f+\phi$

We get the first schema above, which we dub the  $yz$ -schema (in terms of the latin "inputs"  $y, z$  at its left margin), from the original  $xyz$ -schema by pivoting on its starred principal (diagonal) entry. Pivoting on the starred principal entry of the  $yz$ -schema we get the  $z$ -schema above, and then by its starred pivot the null-schema above. Also, by pivoting on the principal entry 2 in the  $z$ -schema, and then rearranging, we get the  $wz$ -schema above.

A solution to our minimum-distance problem can be read from the  $wz$ -schema: set the inputs  $\xi = 0$ ,  $\eta = 0$  and  $w = 0$ ,  $z = 0$  to get  $x = 1$ ,  $y = 2$  and  $\omega = 1$ ,  $\zeta = 2$  and  $f+\phi = 3$ . Also  $f-\phi = \omega w + \xi x + \eta y + \zeta z = 0$ , so  $f = \phi = 3/2$ . The point  $(1, 2, 0)$ , on the edge  $w = 0$ ,  $z = 0$  of the tetrahedron, is the point of the tetrahedron nearest to the given point  $(2, 3, -1)$ . Half the square of this minimal distance is  $3/2$ .

In the  $z$ -schema set the inputs  $\omega = 0$ ,  $\xi = 0$ ,  $\eta = 0$  and  $z = 0$  to get  $w = -2$ ,  $x = 2$ ,  $y = 3$  and  $\zeta = 1$  and  $f+\phi = 1$ . Here again  $f-\phi = \omega w + \xi x + \eta y + \zeta z = 0$ . The point  $(2, 3, 0)$  is the foot of the perpendicular from  $(2, 3, -1)$  to the plane  $z = 0$  bounding the tetrahedron but  $w = -2$  shows that the point  $(2, 3, -1)$  lies on the wrong side of the plane  $w = 0$ . Also, in the  $yz$ -schema set the inputs  $\omega = 0$ ,  $\xi = 0$  and  $y = 0$ ,  $z = 0$  to get  $w = 1$ ,  $x = 2$  and  $\eta = -3$ ,  $\zeta = 1$  and  $f+\phi = 10$ . Again  $f-\phi = \omega w + \xi x + \eta y + \zeta z = 0$ . The point  $(2, 0, 0)$  is on the  $y = 0$ ,  $z = 0$  edge of the tetrahedron and is the foot of the perpendicular on this edge from  $(2, 3, -1)$  but the gradient  $\nabla f$  at  $(2, 0, 0)$  is unfavorably directed due to  $\eta = -3$ .



This diagram depicts the full equivalence-class of  $15(=2^4-1)$  schemata [there is no  $wxyz$ -schema, for otherwise  $w = x = y = z = 0$  would be part of a solution]. Each node represents a schema and each arc a principal pivot step (in either direction). Arrows mark the four pivot steps used above to get from the  $xyz$ -schema to the  $yz$ -,  $z$ -, null-, and  $wz$ -schemata. At the same time the diagram depicts the combinatorial (incidence) structure of the constraint tetrahedron!

## DUAL QUADRATIC PROGRAMS

We seek to minimize  $f$  for nonnegative  $x = (x_1, \dots, x_{m+n})^T$ , the "Latin problem," and to maximize  $\phi$  for nonnegative  $\xi = (\xi_1, \dots, \xi_{m+n})$ , the "Greek problem," subject to the quadratic equation

$$f - \phi = \xi x$$

and to any one of a finite class of equivalent systems of linear equations, each given by a schema

$\xi_1$	$p_{11}$	$\dots$	$p_{1m}$	$a_{11}$	$\dots$	$a_{1n}$	$b_1$
$(\xi) :$	$:$	$(P)$	$:$	$:$	$(A)$	$:$	$:$
$\xi_m$	$p_{m1}$	$\dots$	$p_{mm}$	$a_{m1}$	$\dots$	$a_{mn}$	$b_m$
$x_{m+1}$	$-a_{11}$	$\dots$	$-a_{m1}$	$q_{11}$	$\dots$	$q_{1n}$	$c_1$
$(x) :$	$:$	$(-A^T)$	$:$	$:$	$(Q)$	$:$	$:$
$x_{m+n}$	$-a_{1n}$	$\dots$	$-a_{mn}$	$q_{n1}$	$\dots$	$q_{nn}$	$c_n$
$1$	$-b_1$	$\dots$	$-b_m$	$c_1$	$\dots$	$c_n$	$2d$
	$=x_1$	$\dots$	$=x_m$	$=x_{m+1}$	$\dots$	$=x_{m+n}$	$=f+\phi$
	$(=x)$			$(=\xi)$			

where  $\bar{1}, \dots, \bar{m+n}$  denotes a permutation of the indices  $1, \dots, m+n$ .  $P$  and  $Q$  are positive semidefinite symmetric square submatrices. Note the bisymmetry: the upper left and lower right "quarters" are symmetric, while the lower left is the negative transpose of the upper right.

With each such schema there goes a pair of quadratic programs dual in the sense of Cottle:

$$(G) \text{ maximize } \phi = d + \xi b - \frac{1}{2} \xi P \xi^T - \frac{1}{2} \bar{x}^T Q \bar{x} \text{ for } \xi \geq 0, \bar{x} = \xi A + \bar{x}^T Q + c \geq 0;$$

$$(L) \text{ minimize } f = d + c \bar{x} + \frac{1}{2} \bar{x}^T Q \bar{x} + \frac{1}{2} \xi P \xi^T \text{ for } \bar{x} \geq 0, \bar{x} = P \xi^T - A \bar{x} - b \geq 0.$$

A value of  $\phi$  (or  $f$ ) is termed feasible if it arises from a solution  $\xi, x$  of the schema having  $\xi \geq 0$  (or  $x \geq 0$ ). For any two solutions  $\xi, x$  and  $\xi', x'$

$$f' - \phi \geq \xi (P \xi'^T - A \bar{x}' - b) + (\xi A + \bar{x}^T Q + c) \bar{x}' = \xi x'$$

since  $(\xi' - \xi)P(\xi' - \xi)^T \geq 0$  and  $(\bar{x}' - \bar{x})^T Q(\bar{x}' - \bar{x}) \geq 0$ . If  $\xi \geq 0$  and  $x' \geq 0$ , then  $f' \geq \phi$ . That is, each feasible  $f$  is an upper bound for all feasible  $\phi$  and each feasible  $\phi$  is a lower bound for all feasible  $f$ . Hence a solution of the above schema such that  $\xi \geq 0, x \geq 0$  and  $\xi x = 0$  is optimal for both programs. In this case  $f = \phi = \frac{1}{2}(f+\phi)$ .

The class of schemata equivalent to the above is generated by three types of pivotal exchange: (1) pivoting on a diagonal entry  $p_{ii} \neq 0$ , which decreases  $m$  by one and increases  $n$  by one; (2) pivoting on a diagonal entry  $q_{jj} \neq 0$ , which decreases  $n$  by one and increases  $m$  by one; (3) pivoting on a nonzero skew pair  $a_{ij}$  and  $-a_{ij}$ , which does not change  $m$  and  $n$ . Under such pivoting the schematic bisymmetry and the positive semidefiniteness of the symmetric square submatrices  $P$  and  $Q$  are preserved. Moreover, the submatrices  $P$  and  $Q$  have constant nullities  $m_0 = m - \text{rank } P$  and  $n_0 = n - \text{rank } Q$ . Within this (finite) class of equivalent schemata there must exist at least one schema with  $m = m_0$ , i.e., with  $P = 0$  or with  $P, A, b$  vacuous if  $m_0 = 0$ , yielding a "pure" Latin program:

$$(L_0) \quad \text{minimize } f = d + c\ddot{x} + \frac{1}{2} \ddot{x}^T Q \ddot{x} \text{ for } \ddot{x} \geq 0, -\dot{x} = A\ddot{x} + b \leq 0.$$

This is the classic type of convex quadratic program for which Dorn (and also Dennis) introduced duality. Also, of course, there is at least one schema with  $n = n_0$ , yielding a dual "pure" Greek program:

$$(G_0) \quad \text{maximize } \phi = d + \dot{\xi}b - \frac{1}{2} \dot{\xi}P\dot{\xi}^T \text{ for } \dot{\xi} \geq 0, \ddot{\xi} = \dot{\xi}A + c \geq 0.$$

The two programs become ordinary dual linear programs (in Dantzig's canonical form) if  $m_0 = m$  and  $n_0 = n$ , so that  $P = 0$  and  $Q = 0$ .

Recent work of Dantzig and Cottle, of Lemke, and of Parsons, shows that within the (finite) class of equivalent schemata there must exist a schema with an obvious solution

$$\dot{\xi} = 0, \ddot{x} = 0, \dot{x} = -b \geq 0, \ddot{\xi} = c \geq 0, f = \phi = d$$

or a schema with an obvious infeasibility (viz., a nonpositive column with its bottom entry negative).

NOTE. The author gratefully acknowledges the collaboration of Dr. T. D. Parsons, Princeton University, and of Dr. Philip Wolfe, IBM Research Center.

R. W. Cottle: Quart. Appl. Math. 21 (1963) 237-243; SIAM Jour. 12 (1964) 663-665.  
G. B. Dantzig & R. W. Cottle: see J. Abadie (ed.), NONLIN. PROGR. (No. Holland 1967).

W. S. Dorn: SIAM Jour. 9 (1961) 51-54; Manag. Sci. 9 (1962-63) 171-208.

C. E. Lemke: Manag. Sci. 11 (1965) 681-689.

T. D. Parsons: Ph.D. Thesis (Princeton 1966).

E. L. Stiefel: INTROD. TO NUMER. MATH., Acad. Press 1963, 1-44.

A. W. Tucker: Oper. Res. 5 (1957) 244-257; see R. Graves and P. Wolfe (eds.), REC. ADV. IN MATH. PROGR. (Wiley 1963) 320-347.

P. S. Wolfe: Quart. Appl. Math. 19 (1961) 239-244.

**MATHEMATICAL PROGRAMMING**

by

**RICHARD W. COTTLE**

at the

**American Mathematical Society Summer Seminar**

on the

**Mathematics of the Decision Sciences**

**Stanford University**

**July - August 1967**

COMPLEMENTARY PIVOT THEORY OF MATHEMATICAL PROGRAMMING

by

Richard W. Cottle and George B. Dantzig\*

TECHNICAL REPORT NO. 16

June 5, 1967

PREPARED UNDER THE AUSPICES

OF

NATIONAL SCIENCE FOUNDATION GRANT GP-3739

Gerald J. Lieberman, Project Director

\*Research partially supported by U. S. Army Research Office  
Contract No. DAHCO4-67-C-0028, Office of Naval Research,  
Contract ONR-N-00014-67-A-0112-0011, U. S. Atomic Energy Commission,  
Contract No. AT(04-3)-326 PA #18, and National Science Foundation  
Grant GP 6431.

DEPARTMENT OF OPERATIONS RESEARCH

STANFORD UNIVERSITY

STANFORD, CALIFORNIA

# COMPLEMENTARY PIVOT THEORY OF MATHEMATICAL PROGRAMMING

by

Richard W. Cottle and George B. Dantzig

1. Formulation. Linear programming, quadratic programming, and bimatrix (two-person, non zero-sum) games lead to the consideration of the following Fundamental Problem: Given a real  $p$ -vector  $q$  and a real  $p \times p$  matrix  $M$ , find vectors  $w$  and  $z$  which satisfy the conditions<sup>\*</sup>

$$(1) \quad w = q + Mz, \quad w \geq 0, z \geq 0$$

$$(2) \quad zw = 0$$

The remainder of this section is devoted to an explanation of why this is so. (There are other fields in which this fundamental problem arises -- see for example [6] and [13] -- but we do not treat them here.) Sections 2 and 3 are concerned with constructive procedures for solving the fundamental problem under various assumptions on the data  $q$  and  $M$ .

---

\* In general, capital roman letters denote matrices while vectors are denoted by lower case roman letters. Whether a vector is a row or a column will always be clear from the context, and consequently we dispense with transpose signs on vectors. In (2), for example,  $zw$  represents the scalar product of  $z$ (row) and  $w$ (column). The superscript <sup>T</sup> indicates the transpose of the matrix to which it is affixed.

Consider first linear programs in the symmetric primal-dual form due to J. von Neumann [20].

Primal linear program: Find a vector  $x$  and minimum  $\bar{z}$  such that

$$(3) \quad Ax \geq b, \quad x \geq 0, \quad \bar{z} = cx$$

Dual linear program: Find a vector  $y$  and maximum  $\underline{z}$  such that

$$(4) \quad yA \leq c, \quad y \geq 0, \quad \underline{z} = yb$$

The duality theorem of linear programming [3] states that  $\min \bar{z} = \max \underline{z}$  when the primal and dual systems (3) and (4), respectively, are consistent or -- in mathematical programming parlance -- "feasible." Since

$$\underline{z} = yb \leq yAx \leq cx = \bar{z}$$

for all primal-feasible  $x$  and dual-feasible  $y$ , one seeks such solutions for which

$$(5) \quad yb = cx$$

The inequality constraints of the primal and dual problems can be converted to equivalent systems of equations in non-negative variables through the introduction of non-negative "slack" variables. Jointly, the systems (3) and (4) are equivalent to

$$(6) \quad \begin{aligned} Ax - v &= b, & v &\geq 0, & x &\geq 0 \\ A^T y + u &= c, & u &\geq 0, & y &\geq 0 \end{aligned}$$

and the linear programming problem becomes one of finding vectors  $u, v, x, y$  such that

$$(7) \quad \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} c \\ -b \end{pmatrix} + \begin{pmatrix} 0 & -A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \quad \begin{matrix} u \geq 0, v \geq 0 \\ x \geq 0, y \geq 0 \end{matrix}$$

and ( by (5) )

$$(8) \quad xu + yv = 0$$

The definitions

$$(9) \quad w = \begin{pmatrix} u \\ v \end{pmatrix}, \quad q = \begin{pmatrix} c \\ -b \end{pmatrix}, \quad M = \begin{pmatrix} 0 & -A^T \\ A & 0 \end{pmatrix}, \quad z = \begin{pmatrix} x \\ y \end{pmatrix}$$

establish the correspondence between (1), (2) and (3), (4).

The quadratic programming problem is typically stated in the following manner: Find a vector  $x$  and minimum  $\bar{z}$  such that

$$(10) \quad Ax \geq b, \quad x \geq 0, \quad \bar{z} = cx + \frac{1}{2}xDx$$

In this formulation, the matrix  $D$  may be assumed to be symmetric. The minimand  $\bar{z}$  is a globally convex function of  $x$  if and only if the quadratic form  $x Dx$  (or matrix  $D$ ) is positive semi-definite, and when this is the case, (10) is called the convex quadratic programming problem. It is immediate that when  $D$  is the zero matrix, (10) reduces to the linear program (3). In this sense, the linear programming problem is a special case of the quadratic programming problem.

For any quadratic programming problem (10), define  $u$  and  $v$  by

$$(11) \quad u = Dx - A^T y + c, \quad v = Ax - b$$

A vector  $x^0$  yields minimum  $\bar{z}$  only if there exists a vector  $y^0$  and vectors  $u^0, v^0$  given by (11) for  $x = x^0$  satisfying

$$(12) \quad x^0 \geq 0, u^0 \geq 0, y^0 \geq 0, v^0 \geq 0$$

$$x^0 u^0 = 0, \quad y^0 v^0 = 0$$

These necessary conditions for a minimum in (10) are a direct consequence of a theorem of H. W. Kuhn and A. W. Tucker [14]. It is well known -- and not difficult to prove from first principles -- that (12), known as the Kuhn-Tucker conditions, are also sufficient in the case of convex quadratic programming. By direct substitution, we have for any feasible vector  $x$ ,

$$\begin{aligned} \bar{z} - \bar{z}^0 &= c(x - x^0) + \frac{1}{2}x Dx - \frac{1}{2}x^0 D x^0 \\ &= u^0(x - x^0) + y^0(v - v^0) + \frac{1}{2}(x - x^0)D(x - x^0) \\ &= u^0 x + y^0 v + \frac{1}{2}(x - x^0)D(x - x^0) \geq 0 \end{aligned}$$

which proves the sufficiency of conditions (12) for a minimum in the convex case.

Thus, the problem of solving a quadratic program leads to a search for solution of the system

$$(13) \quad \begin{aligned} u &= Dx - A^T y + c & x &\geq 0, y \geq 0 \\ v &= Ax - b & u &\geq 0, v \geq 0 \end{aligned}$$

$$(14) \quad xu + yv = 0$$

The definitions

$$(15) \quad w = \begin{pmatrix} u \\ v \end{pmatrix}, \quad q = \begin{pmatrix} c \\ -b \end{pmatrix}, \quad M = \begin{pmatrix} D & -A^T \\ A & 0 \end{pmatrix}, \quad z = \begin{pmatrix} x \\ y \end{pmatrix}$$

establish (13), (14) as a problem of the form (1), (2).

Dual of a convex quadratic program. From (15) one is led naturally to the consideration of a matrix  $M = \begin{pmatrix} D & -A^T \\ A & E \end{pmatrix}$  wherein  $E$ , like  $D$ , is positive semi-definite. It is shown in [1] that the

Primal quadratic program: Find  $x$  and minimum  $\bar{z}$  such that

$$(16) \quad Ax + Ey \geq b, \quad x \geq 0, \quad \bar{z} = cx + \frac{1}{2}(xDx + yEy)$$

has the associated

Dual quadratic program: Find  $y$  and maximum  $\underline{z}$  such that

$$(17) \quad -Dx + A^T y \leq c, \quad y \geq 0, \quad \underline{z} = by - \frac{1}{2}(xDx - yEy)$$

All the results of duality in linear programming extend to these

problems, and indeed they are jointly solvable if either is solvable.

When  $E = 0$ , the primal problem is just (10) for which W. S. Dorn

[5] first established the duality theory later extended in [1]. When

both  $D$  and  $E$  are zero matrices, this dual pair (16), (17)

reduces to the dual pair of linear programs (3), (4).

REMARKS. (a) The minimand in (10) is strictly convex if and only if the quadratic form  $x^T D x$  is positive definite. Any feasible strictly convex quadratic program has a unique minimizing solution  $x^0$ .

(b) When  $D$  and  $E$  are positive semi-definite (the case of convex quadratic programming), so is

$$M = \begin{pmatrix} D & -A^T \\ A & E \end{pmatrix}$$

A bimatrix (or two-person nonzero-sum) game,  $\Gamma(A,B)$ , is given by a pair of  $m \times n$  matrices  $A$  and  $B$ . One party, called the row player, has  $m$  pure strategies which are identified with the rows of  $A$ . The other party, called the column player, has  $n$  pure strategies which correspond to the columns of  $B$ . If the row player uses his  $i$ th pure strategy and the column player uses his  $j$ th pure strategy, then their respective losses are defined as  $a_{ij}$  and  $b_{ij}$ , respectively. Using mixed strategies

$$x = (x_1, \dots, x_m) \geq 0, \quad \sum_{i=1}^m x_i = 1$$

$$y = (y_1, \dots, y_n) \geq 0, \quad \sum_{j=1}^n y_j = 1$$

their expected losses are  $x^T A y$  and  $x^T B y$ , respectively. (A component in a mixed strategy is interpreted as the probability with which the player uses the corresponding pure strategy.)

A pair  $(x^0, y^0)$  of mixed strategies is a Nash [19] equilibrium point of  $\Gamma(A,B)$  if

$$x^0 A y^0 \leq x A y^0 \quad \text{all mixed strategies } x$$

$$x^0 B y^0 \leq x^0 B y \quad \text{all mixed strategies } y$$

It is evident (see for example [15]) that if  $(x^0, y^0)$  is an equilibrium point of  $\Gamma(A, B)$ , then it is also an equilibrium point for the game  $\Gamma(A', B')$  in which

$$A' = [a_{ij} + K], \quad B' = [b_{ij} + L]$$

where  $K$  and  $L$  are arbitrary scalars. Hence there is no loss of generality in assuming that  $A > 0$  and  $B > 0$ , and we shall make this assumption hereafter.

Next, by letting  $e_k$  denote the  $k$ -vector all of whose components are unity, it is easily shown that  $(x^0, y^0)$  is an equilibrium point of  $\Gamma(A, B)$  if and only if

$$(18) \quad (x^0 A y^0) e_m \leq A y^0 \quad (A > 0)$$

$$(19) \quad (x^0 B y^0) e_n \leq B^T x^0 \quad (B > 0)$$

This characterization of an equilibrium point leads to a theorem which relates the equilibrium-point problem to a system of the form (1), (2). For  $A > 0$  and  $B > 0$ , if  $u^*, v^*, x^*, y^*$  is a solution of the system

$$(20) \quad u = A y - e_m \quad u \geq 0, \quad y \geq 0$$

$$v = B^T x - e_n \quad v \geq 0, \quad x \geq 0$$

$$(21) \quad xu + yv = 0$$

then

$$(x^0, y^0) = \left( \frac{x^*}{x^* e_m}, \frac{y^*}{y^* e_n} \right)$$

is an equilibrium point of  $\Gamma(A, B)$ . Conversely, if  $(x^0, y^0)$  is an equilibrium point of  $\Gamma(A, B)$  then

$$(x^*, y^*) = \left( \frac{x^0}{x^0 B y^0}, \frac{y^0}{x^0 A y^0} \right)$$

is a solution of (20), (21). The latter system is clearly of the form (1), (2), where

$$w = \begin{pmatrix} u \\ v \end{pmatrix}, \quad q = \begin{pmatrix} -e_m \\ -e_n \end{pmatrix}, \quad M = \begin{pmatrix} 0 & A \\ B^T & 0 \end{pmatrix}, \quad z = \begin{pmatrix} x \\ y \end{pmatrix}$$

Notice that the assumption  $A > 0, B > 0$  precludes the possibility of the matrix  $M$  above belonging to the positive semi-definite class.

The existence of an equilibrium point for  $\Gamma(A, B)$  was established by J. Nash [19] whose proof employs the Brouwer Fixed-Point Theorem. Recently, an elementary constructive proof was discovered by C. E. Lemke and J. T. Howson, Jr. [15].

2. Lemke's iterative solution of the fundamental problem. This section is concerned with the iterative technique of Lemke and Howson for finding equilibrium points of bimatrix games which was later extended by Lemke to the fundamental problem (1), (2). We introduce

first some terminology common to the subject of this section and the next. Consider the system of linear equations

$$(22) \quad w = q + Mz$$

where, for the moment, the  $p$ -vector  $q$  and the  $p \times p$  matrix  $M$  are arbitrary. Both  $w$  and  $z$  are  $p$ -vectors.

For  $i = 1, \dots, p$  the corresponding variables  $z_i$  and  $w_i$  are called complementary and each is the complement of the other. A complementary solution of (22) is a pair of vectors satisfying (22) and

$$(23) \quad z_i w_i = 0, \quad i = 1, \dots, p$$

Notice that a solution  $(w; z)$  of (1), (2) is a nonnegative complementary solution of (22). Finally, a solution of (22) will be called almost-complementary if it satisfies (23) except for one value of  $i$ , say  $i = \beta$ . That is,  $z_\beta \neq 0$ ,  $w_\beta \neq 0$ .

In general, the procedure assumes as given an extreme point of the convex set

$$Z = \left\{ z \mid w = q + Mz \geq 0, z \geq 0 \right\}$$

which also happens to be the endpoint of an almost-complementary ray (unbounded edge) of  $Z$ . Each point of this ray satisfies (23) but for one value of  $i$ , say  $\beta$ . It is not always easy to find such a starting point for an arbitrary  $M$ . Yet there are two important realisations of the fundamental problem which can be so initiated.

The first is the bimatrix game case to be discussed soon; the second is the case where an entire column of  $M$  is positive. The latter property can always be artificially induced by augmenting  $M$  with an additional positive column; as we shall see, this turns out to be a useful device for initiating the procedure with a general  $M$ .

Each iteration corresponds to motion from an extreme point  $P_i$  along an edge of  $Z$  all points of which are almost-complementary solutions of (22). If this edge is bounded, an adjacent extreme point  $P_{i+1}$  is reached which is either complementary or almost-complementary. The process terminates if (i) the edge is unbounded (a ray), (ii)  $P_{i+1}$  is a previously generated extreme point, or (iii)  $P_{i+1}$  is a complementary extreme point.

Under the assumption of nondegeneracy, the extreme points of  $Z$  are in one-to-one correspondence with the basic feasible solutions of (22) (See [3]). Still under this assumption, a complementary basic feasible solution is one in which the complement of each basic variable is nonbasic. The goal is to obtain a basic feasible solution with such a property. In an almost-complementary basic feasible of (23), there will be exactly one index, say  $\beta$ , such that both  $w_\beta$  and  $z_\beta$  are basic variables. Likewise, there will be exactly one index, say  $\nu$ , such that both  $w_\nu$  and  $z_\nu$  are nonbasic variables\*.

---

\* C. van de Panne and A. Whinston [21] have used the appropriate terms basic and nonbasic pair for  $\{w_\beta, z_\beta\}$  and  $\{w_\nu, z_\nu\}$  respectively.

An almost-complementary edge is generated by holding all nonbasic variables at value zero and increasing either  $z_v$  or  $w_v$  of the nonbasic pair  $z_v, w_v$ . There are consequently exactly two almost-complementary edges associated with an almost-complementary extreme point (corresponding to an almost-complementary basic feasible solution).

Suppose that  $z_v$  is the nonbasic variable to be increased. The values of the basic variables will change linearly with the changes in  $z_v$ . For sufficiently small positive values of  $z_v$ , the almost-complementary solution remains feasible. This is a consequence of the nondegeneracy assumption. But in order to retain feasibility, the values of the basic variables must be prevented from becoming negative.

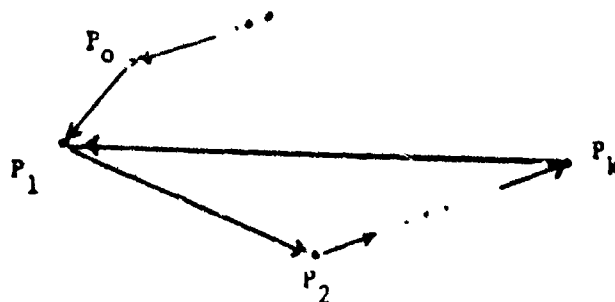
If the value of  $z_v$  can be made arbitrarily large without forcing any basic variable to become negative, then a ray is generated. In this event, the process terminates. However, if some basic variable blocks the increase of  $z_v$  (i.e. vanishes for a positive value of  $z_v$ ), then a new basic solution is obtained which is either complementary or almost-complementary. A complementary solution occurs only if a member of the basic pair blocks  $z_v$ . A new almost-complementary extreme point solution is obtained if the blocking occurs otherwise. In the complementary case, we have the desired result: a complementary basic feasible solution. In the almost-complementary case, the nondegeneracy assumption guarantees the uniqueness of the blocking variable. It will become nonbasic in

place of  $z_v$  and its index becomes the new value of  $v$ .

The complementary rule. The complement of the (now nonbasic) blocking variable -- or equivalently put, the other member of the "new" nonbasic pair -- is the next nonbasic variable to be increased. The procedure consists of the iteration of these steps. The generated sequence of almost-complementary extreme points and edges is called an almost-complementary path.

THEOREM 1. Along an almost-complementary path, the only almost-complementary basic feasible solution which can re-occur is the initial one.

PROOF: We assume that all basic feasible solutions of (22) are nondegenerate. (This can be assured by any of the standard lexicographic techniques [3] for resolving the ambiguities of degeneracy.) Suppose, contrary to the assertion of the theorem, that the procedure generates a sequence of almost-complementary basic feasible solutions in which a term other than the first one ( $P_0$  in the figure below) is repeated (say  $P_1$ ). By the nondegeneracy assumption, the extreme points of  $Z$  are in one-one correspondence with basic feasible solutions of (22). Let  $P_2$  denote the successor of  $P_1$  and let  $P_k$  denote the second predecessor to  $P_1$  namely the one along the path just before the return to  $P_1$ .



The extreme points  $P_c, P_2, P_k$  are distinct and each is adjacent to  $P_1$  along an almost-complementary edge. But there are only two such edges at  $P_1$ . This contradiction completes the proof.

We can immediately state the

COROLLARY. If the almost-complementary path is initiated at the endpoint of an almost-complementary ray, the procedure must terminate either in a different ray or a complementary basic feasible solution.

It is easy to show by examples that starting from an almost-complementary basic feasible solution which is not the endpoint of an almost-complementary ray, the procedure can return to the initial point regardless of the existence or non-existence of a solution to (1), (2).

EXAMPLE 1. The set  $Z$  associated with

$$q = \begin{pmatrix} 1 \\ -1 \\ 3 \end{pmatrix} \quad M = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ -1 & -1 & -1 \end{pmatrix}$$

is nonempty and bounded. It is clear that no solution of (1) can also satisfy (2) since  $z_1 v_1 > 0$ . Let the extreme point corresponding to the solution  $w = (1, 0, 0)$ ,  $z = (1, 0, 2)$  be the initial point of a path which begins by increasing  $z_2$ . This will return to the initial extreme point after 4 iterations.

EXAMPLE 2. The set  $Z$  associated with

$$q = \begin{pmatrix} 1 \\ -1 \\ 3 \\ 1 \end{pmatrix}, \quad M = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ -1 & -1 & -1 & -1 \\ 0 & 0 & 0 & -1 \end{pmatrix}$$

is likewise nonempty and bounded. The corresponding fundamental problem (1), (2) has a complementary solution  $w = (1, 0, 1, 0)$ ,  $z = (0, 1, 0, 1)$ . Yet by starting at  $w = (1, 2, 0, 1)$ ,  $z = (3, 0, 0, 0)$  and increasing  $z_3$ , the method generates a path which returns to its starting point after 4 iterations.

Furthermore, even if the procedure is initiated from an extreme point at the end of an almost-complementary ray, termination in a ray is possible whether or not the fundamental problem has a solution.

EXAMPLE 3. Given the data

$$q = \begin{pmatrix} 1 \\ -1 \\ 3 \\ 1 \end{pmatrix}, \quad M = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & -1 & -1 & 1 \\ 0 & 0 & 0 & -1 \end{pmatrix}$$

the point of  $Z$  corresponding to  $w = (1, 0, 4, 1)$ ,  $z = (1, 0, 0, 0)$  is at the end of an almost-complementary ray,  $w = (1, w_2, 4 + w_2, 1)$ ,  $z = (1 + w_2, 0, 0, 0)$ . Moving along the edge generated by increasing  $z_2$  leads to a new almost-complementary extreme point at which the required increase of  $z_3$  is unblocked, so that the process terminates

in a ray, and yet the fundamental problem is solved by

$$w = (2, 0, 1, 0), \quad z = (0, 1, 0, 1).$$

EXAMPLE 4. In the problem with

$$q = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \quad M = \begin{pmatrix} 0 & 0 \\ 1 & -1 \end{pmatrix}$$

the inequalities (1) have solutions, but none of them satisfy (2).

The point corresponding to  $(w; z) = (1, 0; 1, 0)$  is at the end of an almost-complementary ray  $w = (1, w_2)$ ,  $z = (w_2, 0)$ . When  $z_2$  is increased, it is not blocked, and the process terminates in a ray.

Consequences of termination in a ray. In this geometrical approach to the fundamental problem, it is useful to interpret algebraically the meaning of termination in an almost-complementary ray. This can be achieved by use of a standard result in linear inequality theory [11], [3].

LEMMA. If  $(w^*; z^*)$  is an almost-complementary basic feasible solution of (22), and  $(w^*; z^*)$  is incident to an almost-complementary ray, there exist p-vectors  $w^h, z^h$  such that

$$(24) \quad w^h = Mz^h, \quad w^h \geq 0, \quad z^h \geq 0, \quad z^h \neq 0$$

and points along the almost-complementary ray are of the form

$$(25) \quad (w^* + \lambda w^h, \quad z^* + \lambda z^h) \quad \lambda \geq 0$$

and satisfy

$$(26) \quad (w_1^* + \lambda w_1^h)(z^* + \lambda z_1^h) = 0 \text{ for all } \lambda \geq 0, \text{ and all } i \neq \beta$$

THEOREM 2. If  $M > 0$ , (22) has a complementary basic feasible solution for any vector  $q$ .

PROOF. Select  $w_1, \dots, w_p$  as the basic variables in (22). We may assume that  $q \not\perp 0$  for otherwise  $(w; z) = (q; 0)$  immediately solves the problem. A starting ray of feasible almost-complementary solutions is generated by taking a sufficiently large value of any nonbasic variable, say  $z_1$ . Reduce  $z_1$  toward zero until it reaches a value  $z_1^0 \geq 0$  at which a unique basic variable (assuming non-degeneracy) becomes zero. An extreme point has then been reached.

The procedure has been initiated in the manner described by the corollary above, and consequently the procedure must terminate either in a complementary basic feasible solution or in an almost-complementary ray after some basic feasible solution  $(w; z^*)$  is reached. We now show that the latter cannot happen. For if it does, conditions (24) - (26) of the lemma obtain with  $\beta = 1$ . Since  $M > 0$  and  $z^h \geq 0$ , this implies  $w^h > 0$ . Hence by (26),  $z_1^* = z_1^h = 0$  for all  $i \neq 1$ . Hence the only variables which change with  $\lambda$  are  $z_1$  and the components of  $w$ . Therefore the final generated ray is the same as the initiating ray, which contradicts the corollary.

THEOREM 3. A bimatrix game  $\Gamma(A, B)$  has an extreme equilibrium point.

PROOF. Initiate the algorithm by choosing the smallest positive value of  $x_1$ , say  $x_1^0$ , such that

$$(27) \quad v = -e_n + B_1^T x_1^0 \geq 0$$

where  $B_1^T$  is the first column of  $B^T$ . With

$$v^0 = -e_n + B_1^T x_1^0$$

it follows (assuming nondegeneracy) that  $v^0$  has exactly one zero component, say the  $r$ -th. The ray is generated by choosing as basic variables  $x_1$  and all the slack variables  $u, v$  except for  $v_r$ . The complement of  $v_r$ , namely  $y_r$ , is chosen as the nonbasic variable to increase indefinitely. For sufficiently large values of  $y_r$ , the basic variables are all nonnegative and the ray so generated is complementary except possibly  $x_1 u_1$  might not equal 0. Letting  $y_r$  decrease toward zero, the initial extreme point is obtained for some positive value of  $y_r$ .

If the procedure does not terminate in an equilibrium point, then by the corollary, it terminates in an almost-complementary ray. The latter implies the existence of a class of almost-complementary solutions of the form<sup>\*</sup>

$$(28) \quad \begin{pmatrix} u^* + \lambda u^h \\ v^* + \lambda v^h \end{pmatrix} = \begin{pmatrix} -e_m \\ -e_n \end{pmatrix} + \begin{pmatrix} 0 & A \\ B^T & 0 \end{pmatrix} \begin{pmatrix} x^* + \lambda x^h \\ y^* + \lambda y^h \end{pmatrix}$$

---

\*The notational analogy with the previously studied case  $M > 0$  is obvious.

$$\begin{array}{ll}
 (29) & (u_i^* + \lambda u_i^h)(x_i^* + \lambda x_i^h) = 0 \quad \text{all } i \neq 1 \\
 (30) & (v_j^* + \lambda v_j^h)(y_j^* + \lambda y_j^h) = 0 \quad \text{all } j
 \end{array} \left. \vphantom{\begin{array}{l} (29) \\ (30) \end{array}} \right\} \text{all } \lambda \geq 0$$

Assume first that  $x^h \neq 0$ . Then  $v^h = B^T x^h > 0$ . By (30),  $y_j^* + \lambda y_j^h = 0$  for all  $j$  and all  $\lambda \geq 0$ . But then  $u^* + \lambda u^h = -e_m < 0$ , a contradiction. Assume next that  $y^h \neq 0$  and  $x^h = 0$ . Then  $u^h = A y^h > 0$ . By (29),  $x_i^* = 0$  for all  $i \neq 1$ ; and  $x_1^h = 0$  for all  $i$ . Hence  $v^h = B^T x^h = 0$  and  $v^*$  is the same as  $v$  defined by (27) since  $x_1$  must be at the smallest value in order that  $(u^*, v^*, x^*, y^*)$  be an extreme-point solution. By the nondegeneracy assumption, only  $v_r^* = 0$ , and  $v_j^* > 0$  for all  $j \neq r$ . Hence (30) implies  $y_j^* + \lambda y_j^h = 0$  for all  $j \neq r$ . It is now clear that the postulated terminating ray is the original ray. This furnishes the desired contradiction. The algorithm must terminate in an equilibrium point of the bimatrix game  $\Gamma(A, B)$ .

A modification of almost-complementary basic sets. Consider the system of equations

$$(31) \quad w = q + e_p z_0 + Mz$$

where  $z_0$  represents an "artificial variable" and  $e_p$  is a  $p$ -vector  $(1, \dots, 1)$ . It is clear that (31) always has nonnegative solutions.

A solution of (31) is called almost-complementary if

$z_i w_i = 0$  for  $i = 1, \dots, p$  and is complementary if, in addition,  $z_0 = 0$ . (See [16, p. 685] where a different but equivalent

definition is given.) In this case, let

$$Z_0 = \{ (z_0, z) \mid w = q + e_p z_0 + Mz \geq 0, z_0 \geq 0, z \geq 0 \}$$

We consider the almost-complementary ray generated by sufficiently large  $z_0$ . The variables  $w_1, \dots, w_p$  are initially basic while  $z_0, z_1, \dots, z_p$  are nonbasic variables. For a sufficiently large value of  $z_0$ , say  $z_0^+$ ,

$$w^+ = q + e_p z_0^+ > 0$$

As  $z_0$  decreases toward zero, the basic variables  $w_1$  decrease. An initial extreme point is reached when  $z_0$  attains the minimum value  $z_0^0$  for which  $w = q + e_p z_0 \geq 0$ . If  $z_0^0 = 0$ , then  $q \geq 0$ ; this is the trivial case for which no algorithm is required. If  $z_0^0 > 0$ , some unique basic variable, say  $w_r$  has reached its lower bound 0. Then  $z_0$  becomes a basic variable in place of  $w_r$  and we have  $v = r$ . Next,  $z_r$ , the complement of  $w_r$ , is to be increased.

The remaining steps of the procedure are now identical to those in the preceding algorithm. After a blocking variable becomes basic, its complement is increased until either a basic variable blocks the increase (by attaining its lower bound 0) or else an almost-complementary ray is generated. There are precisely two forms of termination. One is in a ray as just described; the other is in the reduction of  $z_0$  to the value 0 and hence the attainment of a complementary basic feasible solution of (31), i.e. a solution of

(1), (2).

Interest now centers on the meaning of termination in an almost-complementary ray solution of (31). For certain classes of matrices, the process described above terminates in an almost-complementary ray if and only if the original system (1) has no solution. In the remainder of this section, we shall amplify the preceding statement.

If termination in an almost-complementary ray occurs after the process reaches a basic feasible solution  $(w^*; z_0^*, z^*)$  corresponding to an extreme point of  $Z_0$ , then there exists a nonzero vector  $(w^h; z_0^h, z^h)$  such that

$$(32) \quad w^h = e_p z_0^h + Mz^h, \quad (w^h; z_0^h, z^h) \geq 0$$

Moreover for every  $\lambda \geq 0$ ,

(33)

$$(w^* + \lambda w^h) = q + e_p (z_0^* + \lambda z_0^h) + M(z^* + \lambda z^h)$$

and

(34)

$$(w_1^* + \lambda z_1^h)(z_1^* + \lambda z_1^h) = 0 \quad i = 1, \dots, p.$$

The case  $z^h = 0$  is ruled out, for otherwise  $z_0^h > 0$  and then  $w^h > 0$  because  $(w^h; z_0^h, z^h) \neq 0$ . Now if  $w^h > 0$ , (34) implies  $z^* + \lambda z^h = z^* = 0$ . This, in turn, implies that the ray is the original one which is not possible.

Furthermore, it follows from the almost-complementarity of solutions along the ray that

$$(35) \quad z_i^{**} w_i^* = z_i^{*h} w_i^h = z_i^{h*} w_i^* = z_i^{hh} w_i^h = 0 \quad i = 1, \dots, p.$$

The individual equations of the system (32) are of the form

$$(36) \quad w_i^h = z_o^h + (Mz^h)_i \quad i = 1, \dots, p.$$

Multiplication of (36) by  $z_i^h$  leads, via (35), to

$$(37) \quad 0 = z_i^h z_o^h + z_i^h (Mz^h)_i \quad i = 1, \dots, p$$

from which we conclude

THEOREM 4: Termination in a ray implies there exists a nonzero nonnegative vector  $z^h$  such that

$$(38) \quad z_i^h (Mz^h)_i \leq 0 \quad i = 1, \dots, p$$

At this juncture, two large classes of matrices  $M$  will be considered. For the first class, we show that termination in a ray implies the inconsistency of the system (1). For the second class, we will show that termination in a ray cannot occur, so that for this class of matrices, (1), (2) always has a solution regardless of what  $q$  is.

The first class mentioned above was introduced by Lemke [16]. These matrices, which we shall refer to as copositive plus, are required to satisfy the two conditions.

$$(39) \quad uMu \geq 0 \quad \text{for all } u \geq 0$$

$$(40) \quad (M + M^T)u = 0 \quad \text{if} \quad uMu = 0 \quad \text{and} \quad u \geq 0$$

Matrices satisfying conditions (39) alone are known in the literature as copositive (see [18], [12].) To our knowledge, there is no reference other than [16] on copositive matrices satisfying the condition (40). However, the class of such matrices is large and includes

- (i) all strictly copositive matrices, i.e. those for which  $uMu \geq 0$  when  $0 \neq u \geq 0$
- (ii) all positive semi-definite matrices, i.e. those for which  $uMu \geq 0$  for all  $u$ .

Positive matrices are obviously strictly copositive while positive definite matrices are both positive semi-definite and strictly copositive. Furthermore, it is possible to "build" matrices satisfying (39) and (40) out of smaller ones. For example, if  $M_1$  and  $M_2$  are matrices satisfying (39) and (40) then so is the block-diagonal matrix

$$M = \begin{pmatrix} M_1 & 0 \\ 0 & M_2 \end{pmatrix}$$

Moreover, if  $M$  satisfies (39) and (40) and  $S$  is any skew-symmetric matrix (of its order), then  $M + S$  satisfies (39) and (40). Consequently, block matrices such as

$$M = \begin{pmatrix} M_1 & -A^T \\ A & M_2 \end{pmatrix}$$

satisfy (39) and (40) if and only if  $M_1$  and  $M_2$  do too. However, as Lemke [16], [17] has pointed out, the matrices encountered in the bimatrix game problem with  $A > 0$  and  $B > 0$  need not satisfy (40). The Lemke-Howson iterative procedure for bimatrix games was given earlier in this section. If applied to bimatrix games, the modification just given always terminates in a ray after just one iteration, as can be verified by taking any example.

The second class, consisting of matrices having positive principal minors, has been studied by numerous investigators; see for example, [2], [4], [8], [9], [10], [22], [24]. In the case of symmetric matrices, those with positive principal minors are positive definite. But the equivalence breaks down in the non-symmetric situation. Nonsymmetric matrices with positive principal minors need not be positive definite. For example, the matrix

$$\begin{pmatrix} 2 & -7 \\ -1 & 4 \end{pmatrix}$$

has positive principal minors but is indefinite and not copositive. However, positive definite matrices are a subset of those with positive principal minors. (See, e.g. [2].)

We shall make use of the fact that  $w = q + Mz$ ,  $(w; z) \geq 0$  has no solution if there exists a vector  $v$  such that

$$(41) \quad vM \leq 0, \quad vq < 0, \quad v \geq 0$$

for otherwise,  $0 \leq vw = vq + vMz < 0$ , a contradiction. Indeed, it is a consequence of J. Farkas' theorem [7] that (1) has no solution if and only if there exists a solution of (41).

**THEOREM 5.** Let  $M$  be copositive plus. If the iterative procedure terminates in a ray, then (1) has no solution.

**PROOF.** Termination in a ray means that a basic feasible solution  $(w^*; z_0^*, z^*)$  will be reached at which conditions (32) - (34) hold and also

$$(42) \quad 0 = z^h w^h = z^h e_p^h z_0^h + z^h Mz^h$$

Since  $M$  is copositive and  $z^h \geq 0$ , both terms on the right side of (42) are nonnegative, hence both are zero. The scalar  $z_0^h = 0$  because  $z^h e_p^h > 0$ . The vanishing of the quadratic form  $z^h Mz^h$  means

$$Mz^h + M^T z^h = 0$$

But by (32),  $z_0^h = 0$  implies that  $w^h = Mz^h \leq 0$ , whence  $M^T z^h \leq 0$  or, what is the same thing,  $z^h M \leq 0$ . Next, by (35),

$$0 = z^* w^h = z^* Mz^h = z^* (-M^T z^h) = -z^h Mz^*$$

and we obtain again by (35)

$$0 = z^h w^* = z^h q + z^h e_p z_o^* + z^h M z^* = z^h q + z^h e_p z_o^*$$

It follows that  $z^h q < 0$  because  $z^h e_p z_o^* > 0$ . The conditions (1) are therefore inconsistent because  $v = z^h$  satisfies (41).

COROLLARY. If  $M$  is strictly copositive, the process terminates in a complementary basic feasible solution of (31).

PROOF. If not, the proof of theorem 5 would imply the existence of a vector  $z^h$  satisfying  $z^h M z^h = 0$ ,  $0 \neq z^h \geq 0$  which contradicts the strict copositivity of  $M$ .

This corollary clearly generalizes Theorem 1. We now turn to the matrices  $M$  having positive principal minors.

THEOREM 6. If  $M$  has positive principal minors, the process terminates in a complementary basic solution of (31) for any  $q$ .

PROOF. We have seen that termination in a ray implies the existence of a nonzero vector  $z^h$  satisfying the inequalities (38). However, Gale and Nikaido [10, Theorem 2] have shown that matrices with positive principal minors are characterized by the impossibility of this event. Hence termination in a ray is not a possible outcome for problems in which  $M$  has positive principal minors.

We can even improve upon this.

THEOREM 7. If  $M$  has the property that for each of its principal submatrices  $\tilde{M}$ , the system

$$\tilde{M}\tilde{z} \leq 0, \quad 0 \neq \tilde{z} \geq 0$$

has no solution, then the process terminates in a complementary basic solution of (31) for any  $q$ .

PROOF. Suppose the process terminates in a ray. From the solution  $(w^h; z_0^h, z^h)$  of the homogeneous system (32), define the vector  $\tilde{w}^h$  of components of  $w^h$  for which the corresponding component of  $z^* + z^h$  is positive. Then by (34)  $\tilde{w}^h = 0$ . Let  $\tilde{z}^h$  be the vector of corresponding components in  $z^h$ . Clearly  $0 \neq \tilde{z}^h \geq 0$ , since  $0 \neq z^h \geq 0$  and any positive component of  $z^h$  is a positive component of  $\tilde{z}^h$  by definition of  $\tilde{w}^h$ . Let  $\tilde{M}$  be the corresponding principal submatrix of  $M$ . Since  $\tilde{M}$  is a matrix of order  $k \geq 1$  we may write

$$0 = \tilde{w}^h = e_k z_0^h + \tilde{M} \tilde{z}^h$$

Hence

$$\tilde{M} \tilde{z}^h \leq 0, \quad 0 \neq \tilde{z}^h \geq 0$$

which is a contradiction.

3. The principal pivoting method. We shall now describe an algorithm proposed by the authors [4] which predates that of Lemke. It evolved from a quadratic programming algorithm of P. Wolfe [26] who was the first to use a type of complementary rule for pivot choice. Our method is applicable to matrices  $M$  that have positive principal minors (in particular to positive definite matrices) and after a minor modification, to positive semi-definite matrices.

In Lemke's procedure for general  $M$ , an artificial variable  $z_0$  is introduced in order to obtain feasible almost-complementary solutions for the augmented problem. In our approach, only variables of the original problem are used, but these can take on initially

negative as well as non-negative values.

A major cycle of the algorithm is initiated with the complementary basic solution  $(w; z) = (q; 0)$ . If  $q \geq 0$ , the procedure is immediately terminated. If  $q \not\geq 0$ , we may assume (relabeling if necessary) that  $w_1 = q_1 < 0$ . An almost-complementary path is generated by increasing  $z_1$ , the complement of the selected negative basic variable. For points along the path,  $z_i w_i = 0$  for  $i \neq 1$ .

Step I. Increase  $z_1$  until it is blocked by a positive basic variable decreasing to zero or by the negative  $w_1$  increasing to zero.

Step II. Make the blocking variable nonbasic by pivoting its complement into the basic set. The major cycle is terminated if  $w_1$  drops out of the basic set of variables. Otherwise, return to Step I.

It will be shown that during a major cycle  $w_1$  increases to zero. At this point, a new complementary basic solution is obtained. However, the number of basic variables with negative values is at least one less than at the beginning of the major cycle. Since there are at most  $p$  negative basic variables, no more than  $p$  major cycles are required to obtain a complementary feasible solution of (22). The proof depends on certain properties of matrices invariant under principal pivoting.

Principal pivot transform of a matrix. Consider the homogeneous system  $v = Mu$  where  $M$  is a square matrix. Here the variables  $v_1, \dots, v_p$  are basic and expressed in terms of the nonbasic variables

$u_1, \dots, u_p$ . Let any subset of the  $v_i$  be made nonbasic and the corresponding  $u_i$  basic. Relabel the full set of basic variables  $\bar{v}$  and the corresponding nonbasic variables  $\bar{u}$ . Let  $\bar{v} = \bar{M}\bar{u}$  express the new basic variables  $\bar{v}$  in terms of the nonbasic ones. The matrix  $\bar{M}$  is called a principal pivot transform of  $M$ . Of course, this transformation can be carried out only if the principal submatrix of  $M$  corresponding to the set of variables  $z_i$  and  $w_i$  interchanged is nonsingular, and this will be assumed whenever the term is used.

THEOREM 8. (Tucker [24]). If a square matrix  $M$  has positive principal minors, so does every principal pivot transform of  $M$ .

The proof of this theorem is easily obtained inductively by exchanging the roles of one complementary pair and evaluating the resulting principal minors in terms of those of  $M$ .

THEOREM 9. If a matrix  $M$  is positive definite or positive semi-definite so is every principal pivot transform of  $M$ .

PROOF. The original proof given by the authors was along the lines of that for the preceding theorem. P. Wolfe has suggested the following elegant proof. Consider  $v = Mu$ . After the principal pivot transformation, let  $\bar{v} = \bar{M}\bar{u}$ , where  $\bar{u}$  is the new set of nonbasic variables. We wish to show that  $\bar{u}\bar{M}\bar{u} = \bar{u}\bar{v} > 0$  if  $uMu = uv > 0$ . If  $M$  is positive definite, the latter is true if  $u \neq 0$ , and the former must hold because every pair  $(\bar{u}_i, \bar{v}_i)$  is identical with  $(u_i, v_i)$  except possibly in reverse order. Hence  $\sum_1 \bar{u}_i \bar{v}_i = \sum_1 u_i v_i > 0$ . The proof in the semi-definite case replaces the

inequality  $>$  by  $\geq$ .

Validity of the algorithm. The proof given below for  $p = 3$  goes through for general  $p$ . Consider

$$w_1 = q_1 + m_{11}z_1 + m_{12}z_2 + m_{13}z_3$$

$$w_2 = q_2 + m_{21}z_1 + m_{22}z_2 + m_{23}z_3$$

$$w_3 = q_3 + m_{31}z_1 + m_{32}z_2 + m_{33}z_3$$

Suppose that  $M$  has positive principal minors so that the diagonal coefficients are all positive:

$$m_{11} > 0, \quad m_{22} > 0, \quad m_{33} > 0$$

Suppose furthermore that some  $q_i$  is negative, say  $q_1 < 0$ . Then the solution  $(w;z) = (q_1, q_2, q_3; 0, 0, 0)$  is complementary, but not feasible because a particular variable, in this case  $w_1$ , which we refer to as distinguished is negative. We now initiate an almost-complementary path by increasing the complement of the distinguished variable, in this case  $z_1$ , which we call the driving variable. Adjusting the basic variables, we have

$$(w;z)^1 = (q_1 + m_{11}z_1, q_2 + m_{21}z_1, q_3 + m_{31}z_1; 0, 0, 0)$$

Note that the distinguished variable  $w_1$  increases strictly with the increase of the driving variable  $z_1$  because  $m_{11} > 0$ . Assuming nondegeneracy, we can increase  $z_1$  by a positive amount before it is blocked either by  $w_1$  reaching zero or by a basic

variable that was positive and is now turning negative.

In the former case, for some positive value  $z_1^*$  of the driving variable  $z_1$ , we have  $w_1 = q_1 + m_{11}z_1^* = 0$ . The solution

$$(w; z)^2 = (0, q_2 + m_{21}z_1^*, q_3 + m_{31}z_1^*; 0, 0, 0)$$

is complementary and has one less negative component. Pivoting on  $m_{11}$ , replaces  $w_1$  by  $z_1$  as a basic variable. By Theorem 8, the matrix  $\bar{M}$  in the new canonical system relabeled  $\bar{w} = \bar{q} + \bar{M}z$  has positive principal minors, allowing the entire major cycle to be repeated.

In the latter case, we have some other basic variable, say  $w_2 = q_2 + m_{21}z_1$  blocking when  $z_1 = z_1^* > 0$ . Then clearly  $m_{21} < 0$  and  $q_2 > 0$ . In this case,

$$(w; z)^2 = (m_{11}z_1^* + q_1, 0, m_{31}z_1^* + q_3; z_1^*, 0, 0)$$

**THEOREM 10.** If the driving variable is blocked by a basic variable other than its complement, a principal pivot exchanging the blocking variable with its complement will permit the further increase of the driving variable.

**PROOF:** Pivoting on  $m_{22}$  generates the canonical system

$$\begin{aligned} w_1 &= \bar{q}_1 + \bar{m}_{11}z_1 + \bar{m}_{12}w_2 + \bar{m}_{13}z_3 \\ z_2 &= \bar{q}_2 + \bar{m}_{21}z_1 + \bar{m}_{22}w_2 + \bar{m}_{23}z_3 \\ w_3 &= \bar{q}_3 + \bar{m}_{31}z_1 + \bar{m}_{32}w_2 + \bar{m}_{33}z_3 \end{aligned}$$

The solution  $(w; z)^2$  must satisfy the above since it is an equivalent system. Therefore setting  $z_1 = z_1^*, w_2 = 0, z_3 = 0$  yields

$$(w; z)^2 = (q_1 + \bar{m}_{11}z_1^*, 0, q_3 + \bar{m}_{31}z_1^*; z_1^*, 0, 0)$$

i.e., the same almost-complementary solution. Increasing  $z_1$  beyond  $z_1^*$  yields

$$(\bar{q}_1 + \bar{m}_{11}z_1, 0, \bar{q}_3 + \bar{m}_{31}z_1; z_1, 0, 0)$$

which is also almost-complementary. The sign of  $\bar{m}_{21}$  is the reverse of  $m_{21}$ , since  $\bar{m}_{21} = -m_{21}/m_{22} > 0$ . Hence  $z_2$  increases with increasing  $z_1 > z_1^*$ ; i.e., the new basic variable replacing  $w_2$  is not blocking. Since  $\bar{M}$  has positive principal minors,  $\bar{m}_{11} > 0$ . Hence  $w_1$  continues to increase with increasing  $z_1 > z_1^*$ .

THEOREM 11. The number of iterations within a major cycle is finite.

PROOF: There are only finitely many possible bases. No basis can be repeated with a larger value of  $z_1$ . To see this, suppose it did for  $z_1^{**} > z_1^*$ . This would imply that some component of the solution turns negative at  $z_1 = z_1^*$  and yet is nonnegative when  $z_1 = z_1^{**}$ . Since the value of a component is linear in  $z_1$  we have a contradiction.

Paraphrase of the principal pivoting method. Along the almost-complementary path there is only one degree of freedom. In the proof of the validity of the algorithm,  $z_1$  was increasing and  $z_2$  was shown to increase. The same class of solutions can be generated

by regarding  $z_2$  as the driving variable and the other variables as adjusting. Hence within each major cycle, the same almost-complementary path can be generated as follows. The first edge is obtained by using the complement of the distinguished variable as the driving variable. As soon as the driving variable is blocked, the following steps are iterated:

- a) replace the blocking variable by the driving variable and terminate the major cycle if the blocking variable is distinguished; if the blocking variable is not distinguished.
- b) let the complement of the blocking variable be the new driving variable and increase it until a new blocking variable is identified; return to a).

The paraphrase form is used in practice.

**THEOREM 12.** The principal pivoting method terminates in a solution of (1), (2) if  $M$  has positive principal minors (and, in particular, if  $M$  is positive definite).

**PROOF.** We have shown that the completion of a major cycle occurs in a finite number of steps, and each one reduces the total number of variables with negative values. Hence in a finite number of steps, this total is reduced to zero and a solution of the fundamental problem (1), (2) is obtained. Since a positive definite matrix has positive principal minors, the method applies to such matrices.

As indicated earlier, the positive semidefinite case can be handled by using the paraphrase form of the algorithm with a minor modification. The reader will find details in [4].

## REFERENCES

- [1] Cottle, R. W., Symmetric dual quadratic programs, Quart. Appl. Math. 21 (1963), 237-243.
- [2] Cottle, R. W., Nonlinear programs with positively bounded Jacobians, J. SIAM Appl. Math. 14 (1966), 147-158.
- [3] Dantzig, G. B., Linear programming and extensions, Princeton University Press, Princeton, 1963.
- [4] Dantzig, G. B. and Cottle, R. W., Positive (semi-) definite programming, ORC 63-18 (RR), May 1963, Operations Research Center, University of California, Berkeley. Revised in Nonlinear programming (J. Abadie, ed.), North-Holland, Amsterdam, 1967, 55-73.
- [5] Dorn, W. S., Duality in quadratic programming, Quart. Appl. Math. 18 (1960), 155-162.
- [6] Du Val, P., The unloading problem for plane curves, Amer. J. Math. 62 (1940), 307-311.
- [7] Farkas, J., Theorie der einfachen Ungleichungen, J. Reine Angew. Math. 124 (1902), 1-27.
- [8] Fiedler, M. and Ptak, V., On matrices with non-positive off-diagonal elements and positive principal minors, Czech. Math. Journal 12 (1962), 382-400.
- [9] Fiedler, M. and Ptak, V., Some generalizations of positive definiteness and monotonicity, Numerische Math. 9 (1966), 163-172.
- [10] Gale, D. and Nikaido, H., The Jacobian matrix and global univalence of mappings, Math. Ann. 159 (1965), 81-93.

- [11] Goldman, A. J., "Resolution and separation theorems for polyhedral convex sets," in Linear inequalities and related systems, (H. W. Kuhn and A. W. Tucker, eds.) Princeton University Press, Princeton, 1956.
- [12] Hall, M., Jr. and Newman, M., Copositive and completely positive quadratic forms, Proc. Camb. Phil. Soc. 59 (1963), 329-339.
- [13] Kilmister, C. W. and Reeve, J. E., Rational mechanics, American Elsevier, New York, 1966, § 5.4.
- [14] Kuhn, H. W. and Tucker, A. W., "Nonlinear programming" in Second Berkeley symposium on mathematical statistics and probability (J. Neyman, ed.) University of California Press, Berkeley, 1951.
- [15] Lemke, C. E. and Howson, J. T., Jr. Equilibrium points of bimatrix games, J. Soc. Indust. Appl. Math. 12 (1964), 413-423.
- [16] Lemke, C. E., Bimatrix equilibrium points and mathematical programming, Management Sci. 11 (1965), 681-689.
- [17] Lemke, C. E., Private communication.
- [18] Motzkin, T. S., Copositive quadratic forms, Nat. Bur. Standards Report 1818 (1952), 11-12.
- [19] Nash, J. F., Noncooperative games, Ann. Math. 54 (1951), 286-295.
- [20] von Neumann, J., "Discussion of a maximum problem," Collected Works VI (A. H. Taub, ed.), Pergamon Press, New York, 1963.
- [21] van de Panne, C. and Winston, A., A comparison of two methods

- for quadratic programming, Operations Res. 14 (1966), 422-441.
- [22] Parson, T. D., A combinatorial approach to convex quadratic programming, Doctoral Dissertation, Department of Mathematics, Princeton University, May 1966.
- [23] Tucker, A. W., "A combinatorial equivalence of matrices," Proceedings of Symposia in Applied Mathematics, Vol. 10 (R. Bellman and M. Hall, eds.) American Mathematical Society, 1960.
- [24] Tucker, A. W., Principal pivotal transforms of square matrices, SIAM Review 5 (1963), p. 305.
- [25] Tucker, A. W., Pivotal Algebra, Lecture notes (by T. D. Parsons), Department of Mathematics, Princeton University, 1965.
- [26] Wolfe, P., The simplex method for quadratic programming, Econometrica 27 (1959), 382-398.

# THE PRINCIPAL PIVOTING METHOD OF QUADRATIC PROGRAMMING

By

Richard W. Cottle  
Stanford University

## I. BACKGROUND

Quadratic programming is concerned with the study of optimization problems which can be posed in the form\*

$$\begin{aligned} (1) \quad & \text{minimize } Q(x) = c^T x + \frac{1}{2} x^T D x \\ & \text{subject to } Ax \geq b \\ & \quad \quad \quad x \geq 0 \end{aligned}$$

The points (vectors) satisfying the side conditions or constraints of the problem (1) are said to be feasible solutions and collectively they form the constraint set

$$(2) \quad \mathcal{C} = \{x \in R^n \mid Ax \geq b, x \geq 0\}$$

which could, of course, be empty, in which case the problem (1) is said to be infeasible. But empty or not,  $\mathcal{C}$  is always a polyhedral convex set. The convexity of the objective function  $Q(x)$  is quite another matter. It is well known that a quadratic function  $Q$  is convex on  $R^n$  if and only if its "quadratic part"  $\frac{1}{2} x^T D x$  is a positive semi-definite form. If the dimension of  $\mathcal{C}$  is less than  $n$ , then  $Q$  could be convex on  $\mathcal{C}$  without being convex on  $R^n$ . A discussion of this possibility can be found in [6]. When  $Q$  is convex on  $R^n$ ,

---

\* All numerical quantities of this problem are understood to be real numbers. The vector  $x$  represents an  $n$ -tuple of variables whose values are to be determined. The matrix  $A$  is assumed to be of order  $m$  by  $n$ , and without loss of generality, the  $n$ -square matrix  $D$  may be regarded as symmetric. The superscript  $T$  denotes transposition. Vector inequalities are equivalent to componentwise inequalities of the same type.

(1) is called the convex quadratic programming problem. It is a genuine extension of the linear programming problem which corresponds to the case in which  $D$  is the  $n$ -square zero matrix.

As implied above, our study stems from problems of the form (1). This is not restrictive in the class of linearly-constrained quadratic minimization problems. For instance, there is a simple technique for converting problems such as

$$(1') \quad \begin{aligned} \text{minimize } Q(x) &= c^T x + \frac{1}{2} x^T D x \\ \text{subject to } &Ax = b \end{aligned}$$

into the inequality-constrained format (1) without having to double the number of constraints and variables. This is treated in the Appendix.

In (1), we seek a global minimum of  $Q(x)$  subject to  $x \in \mathcal{C}$ , that is, an  $\bar{x} \in \mathcal{C}$  satisfying  $Q(\bar{x}) \leq Q(x)$  for all  $x \in \mathcal{C}$ . Such an  $\bar{x}$  is said to be an optimal solution of the problem. In the usage adopted here, no vector can be optimal if it is not also feasible.

The necessary conditions of optimality for problem (1)—found by applying the celebrated theorem of Kuhn and Tucker [19]—state that if  $\bar{x}$  is an optimal solution to the problem (1), there exists a vector  $\bar{y}$  such that

$$(3) \quad \begin{aligned} c + D\bar{x} - A^T \bar{y} &\geq 0 \\ -b + A\bar{x} &\geq 0 \\ \bar{x} &\geq 0 \\ \bar{y} &\geq 0 \\ \bar{x}^T [c + D\bar{x} - A^T \bar{y}] &= 0 \\ \bar{y}^T [-b + A\bar{x}] &= 0 \end{aligned}$$

Moreover, these conditions are sufficient when (1) is a convex quadratic program.

The duality theory for quadratic programming completely embraces that for linear programming. Thus, when  $D$  is positive semi-definite, the primal problem (1) has the dual

$$\begin{aligned}
 (4) \quad & \text{maximize } F(x,y) = b^T y - \frac{1}{2} x^T D x \\
 & \text{subject to } c + D x - A^T y \geq 0 \\
 & \quad (x \geq 0) \\
 & \quad y \geq 0
 \end{aligned}$$

The duality of the pair (1),(4) was first discovered by Dorn [12] and Dennis [11]. Later, in [3], the author symmetrized the duality theory by amending the primal problem to read

$$\begin{aligned}
 (1') \quad & \text{minimize } Q(x,y) = c^T x + \frac{1}{2} x^T D x + \frac{1}{2} y^T E y \\
 & \text{subject to } -b + A x + E y \geq 0 \\
 & \quad x \geq 0 \\
 & \quad (y \geq 0)
 \end{aligned}$$

where  $E$ , like  $D$ , is symmetric and positive semi-definite. The Kuhn-Tucker conditions for (1') are

$$\begin{aligned}
 (3') \quad & c + D \bar{x} - A^T \bar{y} \geq 0 \\
 & -b + A \bar{x} + E \bar{y} \geq 0 \\
 & \quad \bar{x} \geq 0 \\
 & \quad \bar{y} \geq 0 \\
 & \bar{x}^T [c + D \bar{x} - A^T \bar{y}] = 0 \\
 & \bar{y}^T [-b + A \bar{x} + E \bar{y}] = 0
 \end{aligned}$$

and the dual of (1') is

$$\begin{aligned}
 (4') \quad & \text{maximize } P(x,y) = b^T y - \frac{1}{2} x^T D x - \frac{1}{2} y^T E y \\
 & \text{subject to} \quad c + D x - A^T y \geq 0 \\
 & \quad \quad \quad (x \geq 0) \\
 & \quad \quad \quad y \geq 0
 \end{aligned}$$

The parentheses around sign-restricted variables indicate that these restrictions can be imposed without loss of generality even though they are not required for the validity of the duality theorems.

For any solution  $(x,y)$  of the orthogonality (or "complementary slackness") conditions

$$\begin{aligned}
 x^T [c + D x - A^T y] &= 0 \\
 y^T [-b + A x + E y] &= 0
 \end{aligned}$$

it follows that

$$\begin{aligned}
 Q(x,y) &= c^T x + \frac{1}{2} x^T D x + \frac{1}{2} y^T E y \\
 &= \frac{1}{2} (c^T x + b^T y) \\
 &= b^T y - \frac{1}{2} x^T D x - \frac{1}{2} y^T E y \\
 &= P(x,y)
 \end{aligned}$$

Therefore the value of either quadratic objective function  $P$  or  $Q$  at any solution of (3') is readily calculated by evaluating the linear function  $\frac{1}{2} (c^T x + b^T y)$

As might be imagined, the system (3') plays a central role in the solution of (1') and (4'), a fact stressed by Wolfe [24]. In order to simplify the manipulation of (3'), it has been found advantageous to represent the block matrix

$$\begin{pmatrix} D & -A^T \\ A & E \end{pmatrix}$$

by the single letter  $M$  and the vectors  $\begin{pmatrix} c \\ -b \end{pmatrix}$ ,  $\begin{pmatrix} x \\ y \end{pmatrix}$  by  $q$  and  $z$ , respectively. With these identifications, (3') takes the simpler form

$$(5) \quad \begin{aligned} q + Mz &\geq 0 \\ z &\geq 0 \\ z^T[q + Mz] &= 0 \end{aligned}$$

We shall call this the fundamental problem. It has been recognized, however, that systems of the form (5) can often be solved without relying on the special structure in the identification above. (See Dantzig and Cottle [8] and Lemke [20].) The structural assumptions can be replaced by more general properties of the matrix  $M$ , such as

- (i) positivity of all principal minors
- (ii) positive semi-definiteness

or generalizations thereof. One such generalization is treated by Lemke [20], another by Ingleton [18] and the author [7].

The study of the fundamental problem (5) has been approached in two ways: one existential, the other constructive. As the names suggest, the existential approach is concerned with conditions which imply the existence—and in some cases, the uniqueness—of solutions to the system, whereas the constructive approach concentrates on the development of efficient computational procedures. The two approaches are not completely disjoint, however. For example, the principal pivoting method described below answers the existence question when the data  $M$  and  $q$  are specified and the matrix  $M$  belongs to an allowable class of matrices. It is also true that existential studies allow one to predict the eventual discovery of a constructive treatment of the problem.

## II. PRINCIPAL PIVOTING

Consider a  $p$ -square matrix  $M$  and a  $p$ -vector  $q$ . For any  $p$ -vector  $z$  the expression  $q + Mz$  defines a mapping  $W:R^p \rightarrow R^p$  and we let

$$(6) \quad w = W(z) = q + Mz$$

We think of (6) as a system of  $p$  linear equations in  $2p$  variables, and in the form above, the variables  $z$  are independent while the variables  $w$  are dependent. In the terminology of linear programming [9], the independent variables are nonbasic and the dependent variables are basic.

A solution  $(w,z)$  to equation (6) is said to be nondegenerate if at most  $p$  of the  $2p$  components  $w_1, \dots, w_p, z_1, \dots, z_p$  equal zero.

To pivot in (6) or an equivalent system is to solve for a currently nonbasic variable in terms of the remaining nonbasic variables and one of the basic variables. Thus pivoting exchanges the roles of two variables with respect to membership in the basis. The specification of these variables singles out a particular entry in the matrix of columns corresponding to the nonbasic variables, and this entry is called the pivotal entry. For the operation to be legitimate, it is necessary and sufficient that the pivotal entry be nonzero.

More generally, a block pivot in (6) or an equivalent system consists in solving for a set of  $k$  currently nonbasic variables in terms of the remaining  $p - k$  nonbasic variables and a set of  $k$  basic variables. For this operation to be possible, it is necessary and sufficient for the corresponding pivotal block (submatrix) to be nonsingular. A block pivot in (6) is called a principal pivot if the pivotal block is a principal submatrix of  $M$ .

The variables  $w_i, z_i$  ( $i = 1, \dots, p$ ) are said to be a complementary pair and each is the complement of the other. The set of basic variables is said

to be complementary (almost-complementary) if it contains no (exactly one) complementary pair.

If the system

$$(7) \quad w^* = q^* + M^* z^*$$

is obtained from (6) by a principal pivot, it is possible to rearrange the rows and columns so that  $w_i^*, z_i^*$  is a complementary pair for each  $i = 1, \dots, p$ . For the sake of the discussion below, we assume that this is always done. We call (7) a principal transform of (6). If  $P$  is a permutation matrix of the same order as  $M$ , then the congruent matrix  $P^T M P$  is called a principal rearrangement of  $M$ .

If we let  $\mathcal{P} = \{1, \dots, p\}$  and  $\mathcal{I}, \mathcal{J} \subseteq \mathcal{P}$ , then  $M_{\mathcal{I}, \mathcal{J}}$  denotes the submatrix of  $M$  formed by deleting the the entries except the  $m_{ij}$  for which  $(i, j)$  belongs to  $\mathcal{I} \times \mathcal{J}$ . Thus  $M_{\mathcal{I}, \mathcal{I}}$  is a principal submatrix of  $M$  and its determinant is called a principal minor of  $M$ . It is a standard convention to define the determinant of the empty matrix to be 1.

It is clear that when  $M$  is a  $p$ -square matrix and  $M^*$  is a principal rearrangement of  $M$ , every principal submatrix of  $M^*$  is a principal submatrix of  $M$ . Consequently, principal rearrangement preserves the character of principal minors. Also clear is the fact that if  $M$  is positive semi-definite, so is any of its principal rearrangements.

Using the notation and definitions above, we may now state a result of Tucker [23].

**THEOREM 1.** If  $M$  has positive principal minors and  $M^*$  is a principal transform of  $M$ , then  $M^*$  has positive principal minors.

**PROOF.** Suppose  $M^*$  is obtained from  $M$  by a block pivot on the principal submatrix

$M_{J,J}$ . Then the conclusion follows from another result of Tucker [22, Theorem 3] which implies that for all  $J \subseteq P$

$$(8) \quad \det M_{J,J}^* = \det M_{J \Delta J, J \Delta J} / \det M_{J,J}$$

where  $\Delta$  represents the symmetric difference operation:

$$J \Delta J = (J \cup J) - (J \cap J)$$

On the strength of this theorem, we say that the class of p-square matrices having positive principal minors is invariant under principal pivoting. The formula (8) yields a similar invariance theorem for the class of matrices with nonnegative principal minors.

The class of positive (semi-)definite matrices of a given order is also invariant under principal pivoting.

**THEOREM 2.** If  $M$  is a p-square positive (semi-)definite matrix and  $M^*$  is a principal transform of  $M$ , then  $M^*$  is positive (semi-)definite.

**PROOF.** We may assume that

$$M = \begin{pmatrix} M_{J,J} & M_{J,I} \\ M_{I,J} & M_{I,I} \end{pmatrix}$$

and that  $M^*$  is obtained from  $M$  by a block pivot on the principal submatrix

$M_{J,J}$ . If  $z = (z_J, z_I)$  and  $w = (w_J, w_I)$  are defined conformally, we may write

$$\begin{aligned} w_J &= M_{J,J} z_J + M_{J,I} z_I \\ w_I &= M_{I,J} z_J + M_{I,I} z_I \end{aligned}$$

The quadratic form  $z^T M z$  can then be expressed as

$$z^T M z = z_J^T w_J + z_I^T w_I$$

After the block pivot on  $M_{J,J}$  we obtain

$$\begin{aligned} z_d &= M_{d,d}^{-1} w_d - M_{d,d}^{-1} M_{d,g} z_g \\ w_g &= M_{g,d} M_{d,d}^{-1} w_d + (M_{g,g} - M_{g,d} M_{d,d}^{-1} M_{d,g}) z_g \end{aligned}$$

and hence

$$M^* = \begin{pmatrix} M_{d,d}^{-1} & -M_{d,d}^{-1} M_{d,g} \\ M_{g,d} M_{d,d}^{-1} & M_{g,g} - M_{g,d} M_{d,d}^{-1} M_{d,g} \end{pmatrix}$$

The associated quadratic form is given by

$$w_d^T z_d + z_g^T w_g = z_d^T w_d + z_g^T w_g$$

Therefore, the range of the quadratic form is invariant under principal pivoting.

When  $M$  is positive semi-definite,  $M^*$  must be also as well. If  $M$  is positive definite, the quadratic form  $z^T M z$  is nonnegative for all  $z$  and vanishes only when  $z = (z_d, z_g) = 0$ . Therefore if  $(w_d, z_g)^T M^* (w_d, z_g)$  vanishes, it follows from the relations above that  $(w_d, z_g) = 0$ . Hence  $M^*$  is positive definite.

REMARK. It should be carefully noted that no assumption of symmetry on  $M$  has been used. For the application suggested in Section I, it would be inappropriate to be hampered by such a restriction.

As we shall see, the principal pivoting method for solving (5) relies rather heavily on these invariance theorems. The method consists of a finite sequence of principal (block) pivots. Each such block pivot which is not a "simple" principal pivot amounts to a finite sequence of simple nonprincipal pivots each of which produces an almost-complementary set of basic variables in the current transform of (6). This means that properties of positive principal minors or positive semi-definiteness can temporarily be lost. It will become clear in the sequel that the following facts are helpful.

THEOREM 3. Let  $A = (a_{ij})$  be a 2-square positive semi-definite matrix. If  $a_{11} = 0$ , then  $a_{12} + a_{21} = 0$ .

PROOF. The associated quadratic form is

$$(a_{12} + a_{21})x_1x_2 + a_{22}x_2^2 \geq 0$$

for all  $x_1, x_2$ . If  $a_{12} + a_{21}$  is anything but zero, the inequality cannot hold for all  $x_1, x_2$ .

THEOREM 4. Let  $A = (a_{ij})$  be a 2-square matrix having the properties:

- (i)  $a_{11} \leq 0$ ;
- (ii)  $a_{21} \leq 0$ ;
- (iii)  $a_{11} + a_{21} < 0$ ;
- (iv) if  $a_{11} < 0$ , then

$$A_1 = \begin{pmatrix} -a_{12}/a_{11} & 1/a_{11} \\ (a_{11}a_{22} - a_{12}a_{21})/a_{11} & a_{21}/a_{11} \end{pmatrix}$$

is positive semi-definite;

- (v) if  $a_{21} < 0$ , then

$$A_2 = \begin{pmatrix} a_{11}/a_{21} & (a_{12}a_{21} - a_{11}a_{22})/a_{21} \\ 1/a_{21} & -a_{22}/a_{21} \end{pmatrix}$$

is positive semi-definite.

Then  $A$  must have the properties:

- (vi)  $a_{12} \geq 0$ ;
- (vii)  $a_{22} \geq 0$ ;
- (viii)  $a_{12} + a_{22} > 0$ .

PROOF. The first three properties of  $A$  imply that  $a_{11} < 0$  or  $a_{21} < 0$ —and perhaps both are negative.

CASE I. If  $a_{11} < 0$ , the principal minors of  $A_1$  are nonnegative since it is positive semi-definite.\* In particular,  $a_{12} \geq 0$  and  $a_{22} \geq 0$ . If both were zero, it would follow from Theorem 3 that  $1/a_{11} = 0$  which is absurd. Hence the properties (vi) to (viii) hold.

CASE II. If  $a_{21} < 0$  and  $a_{11} = 0$ , the positive semi-definiteness of  $A_2$  implies  $a_{12} = -1/a_{21} > 0$  and  $a_{22} \geq 0$ . Hence the required conditions hold again.

The theorem can be varied and established even more easily for the case of matrices having positive principal minors. The proof is omitted.

THEOREM 4'. Let  $A = (a_{ij})$  be a 2-square matrix having the properties:

- (i)  $a_{11} < 0$
- (ii)  $a_{21} < 0$
- (iii) the matrix

$$A_1 = \begin{pmatrix} -a_{12}/a_{11} & 1/a_{11} \\ (a_{11}a_{22} - a_{12}a_{21})/a_{11} & a_{21}/a_{11} \end{pmatrix}$$

has positive principal minors;

- (iv) the matrix

$$A_2 = \begin{pmatrix} a_{11}/a_{21} & (a_{12}a_{21} - a_{11}a_{22})/a_{21} \\ 1/a_{21} & -a_{22}/a_{21} \end{pmatrix}$$

has positive principal minors.

Then  $A$  must have the properties:

- (v)  $a_{12} > 0$ ;
- (vi)  $a_{22} > 0$ .

---

\* Positivity of principal minors is a necessary condition of positive semi-definite matrices, regardless of symmetry.

### III. SOME PROPERTIES OF THE FUNDAMENTAL SYSTEM

The linear inequalities

$$(9) \quad \begin{aligned} q + Mz &\geq 0 \\ z &\geq 0 \end{aligned}$$

or equivalently,

$$(9') \quad \begin{aligned} w &= q - Mz \\ w &\geq 0 \\ z &\geq 0 \end{aligned}$$

will be called the fundamental system. For the moment, we will assume only that  $M$  is a  $p$ -square matrix,  $q \in \mathbb{R}^p$ ,  $z \in \mathbb{R}^p$ .

It is a straightforward consequence of Farkas' Theorem [13] that (9) has no solution if, and only if, there exists a  $p$ -vector  $v$  satisfying

$$(10) \quad v^T M \leq 0, \quad v^T q < 0, \quad v \geq 0$$

It has been shown [5] that when  $M$  has positive principal minors, (10) has no solution, and hence (9) is consistent regardless of what  $q$  may be. Moreover, it was shown there that in this case, the fundamental problem always has a solution; the uniqueness of the solution was first pointed out by Ingleton [18] and by the author in [7]. The key ingredient in the proof is the fact, due to Gale and Nikaido [15], that  $M$  has positive principal minors if, and only if,

$$(11) \quad z_1(Mz)_1 \leq 0, \quad 1 = 1, \dots, p \quad \text{implies } z = 0$$

In the positive semi-definite case, the fundamental system need not be consistent, but when it is, the fundamental problem (5) has a solution. If the solution is also nondegenerate, it is unique. See [4] (existence) and [20] (uniqueness).

The principal pivoting method is applicable to fundamental problems (5) in which  $M$  has positive principal minors or is positive semi-definite. It works with solutions of  $w = q + Mz$  rather than with solutions of the full system (9'). Hence no a priori information regarding feasibility need be given. For this reason, the method incorporates a device for recognizing infeasibility when it occurs in the positive semi-definite case. The device rests on Theorem 4 and the following result proved in [10] by Dantzig and the author.

THEOREM 5. Let  $M$  be a  $p$ -square positive semi-definite matrix, and let  $q \in R^p$ .

If for some index  $r$ ,  $1 \leq r \leq p$ ,

$$(i) \quad q_r < 0$$

$$(ii) \quad m_{rr} = 0$$

$$(iii) \quad m_{1r} \geq 0$$

the system (9') has no solution.

This is proved by noting that the hypotheses lead via Theorem 3 to the conclusion that the  $r$ -th equation

$$w_r = q_r + \sum_{j=1}^p m_{rj} z_j$$

can have no nonnegative solution.\*

Although it seems unimportant from the computational standpoint, it is interesting to see that for a large class of matrices, the solutions to (9) must form either an empty or an unbounded set.

---

\* It is worth mentioning that Theorems 3 and 5 are valid for the class of "copositive plus" matrices introduced by Lemke [20, Theorem 4]. However, our Theorem 2 is not valid for this class of matrices, and therein lies a limitation of the principal pivoting method.

THEOREM 6. Let  $M$  be a  $p$ -square matrix and let  $q \in \mathbb{R}^p$ . The set

$$Z(q, M) = \{z \in \mathbb{R}^p \mid q + Mz \geq 0, z \geq 0\}$$

is unbounded if, and only if, it is nonempty and

$$(12) \quad w^T M < 0, \quad w \geq 0$$

has no solution.

PROOF. Suppose  $Z(q, M)$  is nonempty. By a theorem of Goldman [16],  $Z(q, M)$  is unbounded if and only if  $Z(0, M) \neq \{0\}$ . But by a standard alternative theorem (see Gale [14, Theorem 2.10]), this is so if and only if (12) has no solution.

The class for which (12) has no solution includes copositive matrices (i.e., those for which  $z^T M z \geq 0$  for all  $z \geq 0$ ) and adequate matrices (as defined by Ingleton [18]); therefore the class includes all positive semi-definite matrices and all those with positive principal minors.

As an application of Theorem 6, consider the interpretation of (9) as the set of constraints obtained by taking a convex quadratic program and composing its constraints with those of its dual, as done in Section I. If either the primal or the dual constraint set is nonempty, then at least one of them must be unbounded. This generalizes results of Clark [2], Charnes, Cooper, and Thompson [1], and Lemke [20].

#### IV. THE PRINCIPAL PIVOTING METHOD

In this section, we shall present a treatment of the principal pivoting method first proposed by Dantzig and the author [10]. The method is applicable to matrices with positive principal minors (so-called P-matrices) and to those which are positive semi-definite. Since the method can be stated more simply for the class of P-matrices, we begin there and subsequently broaden the discussion to the positive semi-definite case. First though, we make some general remarks.

It is convenient to represent (6) in tabular form as follows:

$$\begin{array}{rcl}
 & & \begin{array}{cccc} 1 & z_1 & \dots & z_p \end{array} \\
 \begin{array}{l} w_1 = \\ w_2 = \\ \vdots \\ w_p = \end{array} & = & \begin{array}{|c|ccc|} \hline q_1 & m_{11} & \dots & m_{1p} \\ \hline q_2 & m_{21} & \dots & m_{2p} \\ \hline \vdots & \vdots & & \vdots \\ \hline q_p & m_{p1} & \dots & m_{pp} \\ \hline \end{array}
 \end{array}$$

The variables to the left of the box are basic and those above it are nonbasic.

The number of negative components in  $q$  is called the index for the fundamental problem (5). If the index is zero, then  $q \geq 0$  and  $z = 0$  solves (5). Obviously (5) has no computational interest unless its index is positive. Then, if we can achieve an equivalent system with an index of zero, the original system is readily solved.

We denote by

$$(6.v) \quad w^{(v)} = q^{(v)} + M^{(v)} z^{(v)}$$

the system obtained from (6.0) after  $v$  iterations. The system (6.0) is just (6).

We say that (6.v) is complementary if  $(w_i^{(v)}, z_i^{(v)}) = \{w_i, z_i\}$ , for  $i = 1, \dots, p$ .

Starting from a complementary system (6.v) with a positive index, we consider

the solution  $(w^{(\nu)}, z^{(\nu)}) = (q^{(\nu)}, 0)$  and select a particular negative component,  $w_r^{(\nu)}$ , which will be called the distinguished variable. Our immediate objective is to make  $w_r^{(\nu)}$  increase to zero without allowing any variable already nonnegative to become negative. In the case of P-matrices which we now consider, this can always be accomplished by a finite number of simple pivots which result in a principal (block) pivot.

Notice that if  $M = M^{(0)}$  is a P-matrix, so is  $M^{(\nu)}$  in any equivalent complementary system  $(6.\nu)$ , i.e., in any principal pivotal transform of  $(6.0)$ . Consequently, the diagonal entries of  $M^{(\nu)}$  are positive, and in particular,

$$\partial w_r^{(\nu)} / \partial z_r^{(\nu)} = m_{rr}^{(\nu)} > 0$$

Verbally, the increase of the nonbasic complement  $z_r^{(\nu)}$  of the distinguished basic variable  $w_r^{(\nu)}$  forces the latter to increase. In this role,  $z_r^{(\nu)}$  is called the driving variable. Increasing the variable  $z_r^{(\nu)}$  to the positive value  $-q_r^{(\nu)} / m_{rr}^{(\nu)}$  drives  $w_r^{(\nu)}$  up to the value zero.

Basis Exit Rule. The driving variable is governed by a rule which states that its increase must stop as soon as a positive basic variable decreases to zero or the distinguished variable increases to zero. The variable which limits the increase of the driving variable is called the blocking variable. In this case, the existence of the blocking variable is clear; nondegeneracy guarantees its uniqueness.

The blocking variable  $w_s^{(\nu)}$  and the driving variable  $z_r^{(\nu)}$  determine a pivotal entry  $m_{sr}^{(\nu)}$  in the tableau corresponding to the basis exchange in which the driving variable replaces the blocking variable. There are two cases. Case A. If  $s = r$ , then the exchange of  $w_r^{(\nu)}$  for  $z_r^{(\nu)}$  is a principal pivot which produces an equivalent complementary system  $(6.\nu+1)$  in which  $M^{(\nu+1)}$  is a P-matrix. Moreover,

(6.2+1) has a lower index than (6.2). If  $\min q_1^{(v+1)} < 0$ , a new distinguished variable is determined and the process is repeated. If  $\min q_1^{(v+1)} \geq 0$ , a solution of the fundamental problem is at hand, viz.,  $(v^{(v+1)}, z^{(v+1)}) = (q^{(v+1)}, 0)$ .

Case B. If  $s \neq r$ , then  $m_{sr}^{(v)} < 0$ , and the exchange of  $v_s^{(v)}$  and  $z_r^{(v)}$  is a non-principal pivot. Before the pivot,

$$(13) \quad \begin{pmatrix} m_{rr}^{(v)} & m_{rs}^{(v)} \\ m_{sr}^{(v)} & m_{ss}^{(v)} \end{pmatrix}$$

is a principal 2-square submatrix of the P-matrix  $M^{(v)}$ . As such, it is a P-

matrix. After the pivot on  $m_{sr}^{(v)}$  this 2-square matrix becomes

$$(14) \quad \begin{pmatrix} m_{rr}^{(v)}/m_{sr}^{(v)} & (m_{rs}^{(v)}m_{sr}^{(v)} - m_{rr}^{(v)}m_{ss}^{(v)})/m_{sr}^{(v)} \\ 1/m_{sr}^{(v)} & -m_{ss}^{(v)}/m_{sr}^{(v)} \end{pmatrix}$$

so which Theorem 4' applies, although in this case, its conclusion

$$(15) \quad (m_{rs}^{(v)}m_{sr}^{(v)} - m_{rr}^{(v)}m_{ss}^{(v)})/m_{sr}^{(v)} > 0, \quad -m_{ss}^{(v)}/m_{sr}^{(v)} > 0$$

is obvious from first principles. The pivot has (i) left the distinguished variable basic at a negative value, and (ii) made the driving variable basic and the blocking variable nonbasic. Now, the system is almost-complementary since the distinguished variable and its complement are basic\* while the blocking variable and its complement are nonbasic.\*\*

One salient feature of the 2-square matrix (14) is that its rows correspond to the basic pair and its columns correspond to the nonbasic pair in the current tableau. We shall call it the pair matrix. The pair matrix is defined only for an almost-complementary system.

---

\* The basic pair.

\* The nonbasic pair.

Basis Entry Rule. The next variable to enter the basis (i.e., the next driving variable) is the complement of the blocking variable which just became nonbasis. Its increase is also governed by the basis exit rule above. Notice that (15) implies that both members of the basic pair will increase with increases in the new driving variable. Hence the existence of a blocking variable is assured. Indeed, when  $M$  is a P-matrix, the distinguished variable is always potentially a blocking variable. If the distinguished variable is not actually the blocking variable at a particular iteration, the pivotal entry is again negative. By the algebra of pivoting, it follows that a new pair matrix is obtained to which Theorem 4' applies.\* Part of the applicability of Theorem 4' stems from the invariance of P-matrices under principal pivoting. The constant applicability of Theorem 4' accounts for the fact that both members of the basic pair will increase as the current driving variable increases. Since nondegeneracy implies that strictly positive increases of the driving variables are always allowed, the procedure must drive the distinguished variable up to zero at which time it is the blocking variable; the corresponding pivot restores the complementarity of the system.

As in linear programming, the sequence of steps by which the distinguished variable is driven to zero is finite. For there are only finitely many basic solutions, each of which corresponds to a unique set of values of the basic variables. Since the distinguished variable and its complement increase strictly from one iteration to the next, no basis can be repeated. The finiteness of the overall procedure now follows from the fact that the index never increases.

\* In this application, the matrices  $A_1$  and  $A_2$  are interpreted as pivotal trans- of the pair matrix which would make them principal submatrices of a P-matrix.

Modifications for the positive semi-definite case. When the matrix  $M$  in the fundamental problem (5) is positive semi-definite, its principal minors are nonnegative rather than strictly positive. This causes certain complications which call for special handling.

Indeed, the system (6) need not have a nonnegative solution in this case. In the modified procedure, the absence of an appropriate pivotal element or, equivalently, the existence of an unblocked driving variable detects this possibility. This is accomplished with Theorems 5, 4, and 2. However, it is necessary to incorporate an artifice to handle the situation in which the driving variable:

- (i) has no effect on the negative distinguished variable;
- (ii) makes at least one other negative basis variable decrease;
- (iii) makes no positive basic variable decrease.

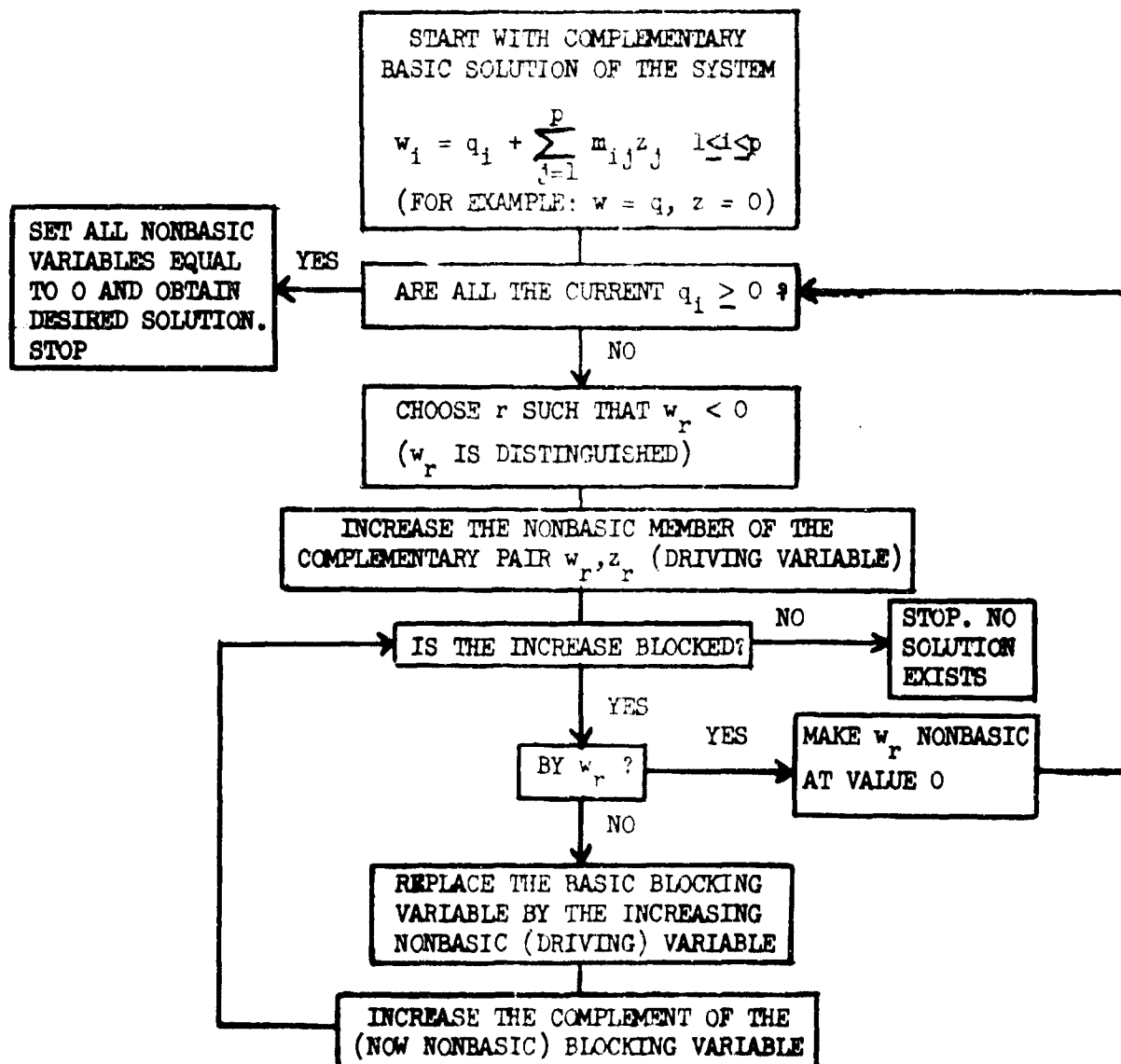
Without modification these conditions would signal an unblocked driving variable, and we want this situation to indicate that no solution to the fundamental problem exists. The artifice we use is to impose a lower bound  $\beta < \min q_i^{(0)}$  on all negative basic variables. A negative variable can therefore block the driving by decreasing to its lower bound; if this happens, the variable becomes nonbasic at its lower bound value,  $\beta$ . However, once a variable becomes nonnegative, zero is its lower bound.

This modification necessitates a change in the notion of basic solution. A basic solution now is one in which the nonbasic variables are set at their current lower bound values, either 0 or  $\beta$ . A solution is nondegenerate if at most  $p$  of the  $2p$  variables of the problem equal 0 or  $\beta$ . Again the nondegeneracy of all solutions is assumed.

The basis entry and exit rules for the positive semi-definite case are as stated above. However, it could happen that the distinguished variable is also the nonbasic driving variable, and in this case, it could be self-blocking.

The principal pivoting method for both of the cases discussed above can be summarized by the following diagram.

FLOW DIAGRAM FOR THE PRINCIPAL PIVOTING METHOD OF QUADRATIC PROGRAMMING



As before, the procedure does not allow nonnegative variables to become negative. Each return to a test of the current "q-column" corresponds to a complementary system with lower index than its predecessor.\* Hence only finitely many returns are possible.

It remains to show that after each nonterminating q-test, a return to the q-test will occur after finitely many steps unless the driving variable is unblocked, in which case the problem cannot be solved due to the infeasibility of (9). As stated earlier, these facts are attributable to the nondegeneracy assumption and Theorems 5, 4, and 2.

Suppose we start from a complementary system. If the driving variable is unblocked, it cannot also be the distinguished variable, since a distinguished driving variable must be a negative nonbasic variable which would not be increased beyond zero. Therefore an unblocked driving variable in a complementary system has a negative basic complement upon which its increase has no effect. Being unblocked, the driving variable must correspond to a nonnegative column in the current tableau. Hence conditions (i), (ii), and (iii) of Theorem 5 hold and the system has no nonnegative solution.

If the driving variable is blocked but not by itself or its complement, the first system after the pivot is almost-complementary and contains a pair matrix. Whenever an almost-complementary system is obtained, the most recent pivotal element is negative. The invariance theorems guarantee the required properties on the matrices  $A_1$ ,  $A_2$  in Theorems 4 and 4'.

---

\* It was pointed out by T.D. Parsons in a private communication that it is possible to return to the q-test and find that column nonnegative while some of the nonbasic variables are at negative values. In such a case, setting all basic variables equal to 0 solves the problem.

If the driving variable in an almost-complementary system is unblocked, its corresponding column in the current tableau must be nonnegative (for otherwise some basic variable would block it). Since the distinguished variable is not blocking, the entry of the pair matrix corresponding to the driving variable and the distinguished variable is zero,<sup>\*</sup> and the other entry in that column of the pair matrix is positive. Pivoting on the latter restores complementarity to the system. After suitable reordering of the columns, it is possible to apply Theorem 5 and declare infeasibility.

Each iteration of the method increases the sum of the distinguished variable and its complement, because the nondegeneracy assumption implies that the driving variable can always be increased. Since there are only finitely many bases and finitely many basic solutions corresponding to each, it is impossible to return to a previously encountered basic solution, and therefore only finitely many steps are required to detect infeasibility or produce a complementary system with lower index than the previous one.

---

\* This could not occur in the case of P-matrices.

# V. A COMPARISON WITH LEMKE'S METHOD

The principal pivoting method invites comparison with the very interesting approach of Lemke [20]. Lemke's method can be viewed as a sequence of almost-complementary pivots resulting in one grand principal pivot which is completely determined after the fundamental problem (5) is embedded in a larger one:

$$(16) \quad \begin{pmatrix} w_0 \\ w \end{pmatrix} = \begin{pmatrix} q_0 \\ q \end{pmatrix} + \begin{pmatrix} 0 & -e^T \\ e & M \end{pmatrix} \begin{pmatrix} z_0 \\ z \end{pmatrix}$$

$$\begin{pmatrix} w_0 \\ w \end{pmatrix} \geq 0, \quad \begin{pmatrix} z_0 \\ z \end{pmatrix} \geq 0$$

$$z_0 w_0 + z^T w = 0$$

where  $e^T = (1, \dots, 1)$ ,  $q_0$  is a suitably large scalar, and  $w_0, z_0$  is a pair of complementary artificial variables. A solution of (16) in which  $z_0 = 0$  is clearly a solution of the original fundamental problem (5).

Lemke has shown that (16) has a solution for any  $p$ -square matrix  $M$ . The question then becomes the significance of a solution to (5) in which  $z_0 > 0$  and hence  $w_0 = 0$ . For a large class of matrices which includes the positive semi-definite class, the answer is that the fundamental problem (5) has no feasible solution. The author and Dantzig [8] have shown that (16) has no solution with  $z_0 > 0$  when  $M$  is a  $P$ -matrix. Hence Lemke's method can always solve (5) when  $M$  is a  $P$ -matrix.

As mentioned in an earlier section, the principal pivoting method is not applicable to the entire class of copositive plus matrices introduced by Lemke since that class is not invariant under principal pivoting. To the author's knowledge, the two methods have never been systematically compared on data to

which they are both applicable.

The Lemke procedure begins with a nonprincipal pivot in the  $z_0$  column. All subsequent systems except the last are almost-complementary and contain the basic pair  $w_0, z_0$ . When  $M$  is positive semi-definite, so is

$$\begin{pmatrix} 0 & -e^T \\ e & M \end{pmatrix}$$

and Theorem 4 can be applied to show that  $w_0$  and  $z_0$  are nonincreasing while  $w_0 + z_0$  is strictly decreasing with increases of successive driving variables. However, this result does not hold for arbitrary matrices.

## VI. APPENDIX

It is well known that an equation is equivalent to a pair of inequalities. Therefore, the system of linear equations

$$(17) \quad Ax = b$$

can always be written as

$$(18) \quad \begin{aligned} Ax &\geq b \\ -Ax &\geq -b \end{aligned}$$

This doubles the number of linear constraints satisfied by  $x$ . If a system of linear inequalities equivalent to (17) is desired, a smaller system than (18) can be used. (Actually, the system will be smaller only when  $m > 1$ .) Since (17) is just

$$(17') \quad \sum_{j=1}^n a_{ij}x_j = b_i \quad i = 1, \dots, m$$

it is equivalent to

$$(19) \quad \sum_{j=1}^n a_{ij}x_j \geq b_i \quad i = 1, \dots, m$$

$$\sum_{j=1}^n \left( -\sum_{i=1}^m a_{ij} \right) x_j \geq -\sum_{i=1}^m b_i$$

which is a system of  $m + 1$  linear inequalities.

Another often-used fact is that any real number  $\xi$  can be represented as the difference of two nonnegative real numbers:

$$\xi = \xi' - \xi'', \quad \xi' \geq 0, \quad \xi'' \geq 0$$

Thus, variables which are not sign restricted can be represented by the difference of two nonnegative variables. In a problem such as (1') where all the variables are unrestricted in sign, this device would double the number of variables. This duplication is also unnecessary. It suffices to write

$$x_j = x'_j - x''_j, \quad x'_j \geq 0, \quad x''_j \geq 0, \quad j = 1, \dots, n$$

This increases the number of variables by 1.

# REFERENCES

- [1] Charnes, A., W.W. Cooper, and G.L. Thompson, Some properties of redundant constraints and extraneous variables in direct and dual linear programming problems, Operations Res. 10 (1962), 711-723.
- [2] Clark, F.E., Remark on the constraint sets in linear programming, Amer. Math. Monthly 68 (1961), 351-352.
- [3] Cottle, R.W., Symmetric dual quadratic programs, Quart. Appl. Math. 21 (1963), 237-243.
- [4] Cottle, R.W., Note on a fundamental theorem in quadratic programming, J. Soc. Indust. Appl. Math. 12 (1964), 663-665.
- [5] Cottle, R.W., Nonlinear programs with positively bounded Jacobians, J. Soc. Indust. Appl. Math. 14 (1966), 147-158.
- [6] Cottle, R.W., On the convexity of quadratic forms over convex sets, Operations Res. 15 (1967), 170-172.
- [7] Cottle, R.W., On a problem in linear inequalities, to appear in J. London Math. Soc. (probably in 1967).
- [8] Cottle, R.W., and G.B. Dantzig, Complementary pivot theory of mathematical programming, Technical Report No. 67-2, Stanford University, April 1967.
- [9] Dantzig, G.B., Linear programming and extensions, Princeton University Press, Princeton, New Jersey, 1963.
- [10] Dantzig, G.B., and R.W. Cottle, "Positive (semi-)definite programming," in (J. Abadie, ed.) Nonlinear programming, North-Holland Publishing Co., Amsterdam, 1967, 55-73.
- [11] Dennis, J.B., Mathematical programming and electrical networks, John Wiley and Sons, New York, 1959.
- [12] Dorn, W.S., Duality in quadratic programming, Quart. Appl. Math. 18 (1960), 155-162.
- [13] Farkas, J., Theorie der einfachen Ungleichungen, J. reine angew. Math. 124 (1902), 1-27.
- [14] Gale, D., The theory of linear economic models, McGraw-Hill, New York, 1960.

- [15] Gale, D., and H. Nikaido, The Jacobian matrix and global univalence of mappings, Math. Ann. 159 (1965), 81-93.
- [16] Goldman, A.J., "Resolution and separation theorems for polyhedral convex sets," in (H.W. Kuhn and A.W. Tucker, eds.) Linear inequalities and related systems, Princeton University Press, Princeton, New Jersey, 1956.
- [17] Graves, R.L., A principal pivoting simplex algorithm for linear and quadratic programming, Operations Res. 15 (1967), 482-494.
- [18] Ingleton, A.W., A problem in linear inequalities, Proc. London Math. Soc. 16 (1966), 519-536.
- [19] Kuhn, H.W., and A.W. Tucker, "Nonlinear programming," in (J. Neyman, ed.) Second Berkeley symposium on mathematical statistics and probability, University of California Press, Berkeley, 1951.
- [20] Lemke, C.E., Bimatrix equilibrium points and mathematical programming, Management Sci. 11 (1965), 681-689.
- [21] Parsons, T.D., A combinatorial approach to convex quadratic programming, Doctoral Dissertation, Department of Mathematics, Princeton University, May 1966.
- [22] Tucker, A.W., "A combinatorial equivalence of matrices," in (R. Bellman and M. Hall, eds.) Proceedings of symposia in applied mathematics, Vol. 10, 1960.
- [23] Tucker, A.W., Principal pivotal transforms of square matrices, SIAM Review 5 (1963), p. 305.
- [24] Wolfe, P., The simplex method for quadratic programming, Econometrica 27 (1959), 382-398.

**NETWORKS AND GRAPHS**

by

**D. R. FULKERSON**

at the

**American Mathematical Society Summer Seminar**

on the

**Mathematics of the Decision Sciences**

**Stanford University**

**July - August 1967**

**MEMORANDUM**

**RM-5368-PR**

**MAY 1967**

**NETWORKS, FRAMES, BLOCKING SYSTEMS**

**D. R. Fulkerson**

This research is supported by the United States Air Force under Project RAND—Contract No. F44620-67-C-0015—monitored by the Directorate of Operational Requirements and Development Plans, Deputy Chief of Staff, Research and Development, Hq USAF. Views or conclusions contained in this Memorandum should not be interpreted as representing the official opinion or policy of the United States Air Force.

**DISTRIBUTION STATEMENT**

Distribution of this document is unlimited.

PREFACE

In this Memorandum, some basic problems concerning flow networks are surveyed and extended to two more general structures: frames of real vector spaces and blocking systems.

### SUMMARY

This paper surveys some basic problems, theorems and constructions for flow networks, and shows how these can be extended to more general combinatorial structures.

One of the generalizations can be roughly described as that obtained by replacing the vertex-edge incidence matrix of an oriented network by an arbitrary real matrix. This leads to the notion of a frame of a subspace of Euclidean  $n$ -space, a concept very closely allied to that of a real matric matroid. Our treatment relates matroid theory and linear programming theory, and thus provides another viewpoint on linear programming, and in particular, on digraphoid-programming.

In the last part of the paper a very general combinatorial structure called a blocking system is given an axiomatic formulation. These systems have arisen in a variety of contexts, including multi-person game theory and abstract covering problems. It is shown that one of the network theorems surveyed in the first part of the paper extends to all blocking systems, and indeed characterizes such systems.

CONTENTS

PREFACE.....	iii
SUMMARY.....	v
INTRODUCTION.....	1
PART I	
1. MAXIMUM FLOW.....	3
2. MINIMUM PATH.....	14
3. MAXIMUM CAPACITY PATH.....	19
4. LENGTH-WIDTH INEQUALITY.....	22
PART II	
1. FRAMES OF REAL SUBSPACES.....	26
2. MATROIDS.....	29
3. GENERALIZED FLOWS AND CUTS.....	33
PART III	
1. AXIOMS AND EXAMPLES.....	46
2. THE MIN-MAX THEOREM.....	51
3. THE LENGTH-WIDTH INEQUALITY AND MAX-FLOW MIN-CUT EQUALITY.....	53
BIBLIOGRAPHY.....	57

## NETWORKS, FRAMES, BLOCKING SYSTEMS

### INTRODUCTION

In this paper we survey a few basic problems, theorems, and constructions concerning flow networks, and describe how some of these can be extended to more general structures.

The paper is divided into three parts.

Most of the material of Part I, which deals with networks, can be found in Ford and Fulkerson [8], or in earlier papers by the same authors. In the main, we limit the discussion in Part I to four network problems: maximum flow, minimum path, maximum capacity path, and the length-width inequality.

Part II extends this discussion to arbitrary real matrices by making use of what we call the frame of a subspace of Euclidean  $n$ -space, a notion very closely related to that of a real matric matroid. In particular, Part II can be specialized to a subclass of real matric matroids introduced and studied by Tutte [21], and called by him regular matroids. Regular matroids have been recently re-investigated by Minty [24], who has given another system of axioms for a dual pair of regular matroids. The resulting structure is called a digraphoid in [24], where it is shown that some of the main theorems of network-programming generalize to digraphoid-programming. Our treatment provides another viewpoint on digraphoid-programming, and indeed on

linear programming in general. It is shown in Part II that the main theorems of Part I have direct analogues for arbitrary real matrices. We want to emphasize, however, that the special network algorithms of Part I do not, so far as we know, have such analogues. Even for the case of digraphoid-programming, we know of nothing better computationally than the simplex method of Dantzig [3]. While the simplex method has proved to be a powerful tool, both theoretically and computationally, it is not yet known whether it is a good algorithm, in the technical sense stressed by Edmonds [6], whereas the network algorithms of Part I are good in this sense.

In Part III a very general combinatorial structure, which we call a blocking system, is given an axiomatic formulation. These systems have arisen previously in a variety of contexts, including multi-person game theory [29] and abstract covering problems [14, 21, 22]. They have recently been studied by Lehman [22], who has given conditions on a blocking system in order that a max-flow min-cut equality or a length-width inequality hold, and also by Edmonds and Fulkerson [7], who have shown that one of the network theorems of Part I extends to all blocking systems, and indeed characterizes such systems.

## PART I. NETWORKS

### 1. MAXIMUM FLOW

Let  $G$  be a graph with edge set  $E$  and vertex set  $V$ . Both  $E$  and  $V$  are assumed finite. The two ends of an edge may be distinct vertices or the same vertex; in the latter case the edge is frequently called a loop. We also allow multiple edges joining the same pair of vertices, or multiple loops on the same vertex.

It will be convenient in this section to orient  $G$  by distinguishing one end of each edge as positive and the other as negative. For a loop these coincide. If  $e \in E$  has positive end  $u \in V$ , negative end  $v \in V$ , we sometimes write  $e = (u, v)$ . For each edge  $e \in E$  and vertex  $v \in V$  we define an integer  $a(v, e)$  as follows. If  $v$  and  $e$  are not incident, or if  $e$  is a loop, then  $a(v, e) = 0$ . Otherwise  $a(v, e) = 1$  or  $-1$  according as  $v$  is the positive or negative end of  $e$ . We call the resulting matrix the vertex-edge incidence matrix of  $G$ .

Suppose now that each edge  $e \in E$  has associated with it a nonnegative real number  $c(e)$ , the capacity of  $e$ . Let  $s$  and  $t$  be two distinguished vertices of  $G$ . A (feasible) flow, of magnitude (or amount)  $\alpha$ , from  $s$  to  $t$  in  $G$  is a real-valued function  $x$  with domain  $E$  that satisfies the linear equations and inequalities

$$(1.1) \quad \sum_{e \in E} a(v, e)x(e) = \begin{cases} \alpha, & v = s, \\ -\alpha, & v = t, \\ 0, & v \neq s, t \end{cases}$$

$$(1.2) \quad -c(e) \leq x(e) \leq c(e), \quad e \in E.$$

Thus  $|x(e)|$  can be thought of as the magnitude of flow in edge  $e$ ; if  $x(e) > 0$ , the direction of flow in  $e$  agrees with the orientation of  $e$ ; if  $x(e) < 0$ , the direction of flow is against the orientation of  $e$ . The equations (1.1) stipulate that  $\alpha$  units of flow leave  $s$  and enter  $t$ , flow being conserved at all other vertices. We call  $s$  the source,  $t$  the sink. The maximum flow problem is that of constructing an  $x$  that satisfies (1.1), (1.2), and maximizes  $\alpha$ .

We can get rid of the asymmetry in equations (1.1) by adding a special edge  $e'$  to  $G$  joining  $s$  and  $t$ , say  $e' \sim (t, s)$ , which returns  $\alpha$  units of flow to  $s$  from  $t$ ; we may take  $c(e')$  large. In other words, by distinguishing one edge  $e'$  of a graph, the maximum flow problem may be viewed as that of maximizing  $x(e')$  subject to (1.2) and the conservation equations

$$(1.1') \quad \sum_{e \in E} a(v, e)x(e) = 0, \quad v \in V.$$

For the moment, we shall continue to work with (1.1) and (1.2), however.

We refer to the graph  $G$  with capacity function  $c$  and distinguished vertex pair  $s, t$  as a (two-terminal) flow network, or briefly, a network. In general, we use the word network in this paper to mean a graph together with one or more real-valued functions defined on its edges.

To state the fundamental theorem about maximum network flow, we require one other notion about graphs, that of a cut. A cut  $K \subset E$  separating  $s$  and  $t$  in a graph  $G$  is a subset of edges that has some edge in common with each path joining  $s$  and  $t$  in  $G$ . We say that  $K$  blocks all such paths. (Here a path joining  $s$  and  $t$  is a sequence of distinct end-to-end edges that starts at  $s$  and ends at  $t$ . Edges may be traversed with or against their orientations in going from  $s$  to  $t$  along the path.) If all edges of  $K$  are deleted from  $G$ , the vertices  $s$  and  $t$  fall in separate components of the new graph. It is intuitively clear that  $\alpha$  in (1.1) is bounded above by

$$(1.3) \quad c(K) = \sum_{e \in K} c(e),$$

the capacity of cut  $K$ . We can prove this from (1.1) and (1.2) by adding those equations of (1.1) corresponding to vertices in the  $s$ -component of the graph  $G'$  gotten from  $G$  by deleting edges of  $K$ . The result is

$$(1.4) \quad \alpha = \sum_{e \in K^+} x(e) - \sum_{e \in K^-} x(e) \leq c(K),$$

where  $K^+$  ( $K^-$ ) consists of those edges of  $K$  with positive (negative) end in the  $s$ -component of  $G'$  and negative (positive) end outside this component. In words, for an arbitrary flow from  $s$  to  $t$  of magnitude  $\alpha$  and an arbitrary

cut separating  $s$  and  $t$ , the net flow across the cut is  $\alpha$ , which is consequently bounded above by the cut capacity. Theorem 1.1 below asserts that equality holds in (1.4) for some flow and some cut, and hence the flow is a maximum flow, the cut a minimum cut [9].

Theorem 1.1. For any network the maximum amount of flow from source to sink is equal to the minimum capacity of all cuts separating source and sink.

Theorem 1.1, the max-flow min-cut theorem, is a combinatorial version, for the special case of the maximum flow problem, of the duality theorem for linear programs, and can be deduced from it [4]. Such a proof makes crucial use of the fact that the vertex-edge incidence matrix of an oriented graph  $G$  is totally unimodular, i.e., every square submatrix has determinant 0 or  $\pm 1$ . A simpler proof of Theorem 1.1 is the second proof given by Ford and Fulkerson [10]. This proof also leads to an efficient algorithm for constructing a maximum flow.

Proof of Theorem 1.1: It suffices to establish the existence of a flow  $x$  and a cut  $K$  for which equality holds in (1.4). Let  $x$  be a maximum flow, of amount  $\alpha$ , from  $s$  to  $t$ . Define a set  $U \subset V$  recursively as follows:

$$(1.5a) \quad s \in U;$$

(1.5b) if  $u \in U$  and  $e \sim (u, v)$  is an edge such that  $x(e) < c(e)$ , then  $v \in U$ ; if  $u \in U$  and  $e \sim (v, u)$  is an edge such that  $x(e) > -c(e)$ , then  $v \in U$ .

We assert that  $t \in \bar{U} = V - U$ . For suppost not. It then follows from the recursive definition of  $U$  that there is a path  $P$  from  $s$  to  $t$  such that  $x(e) < c(e)$  on edges  $e \in P^+$  and  $x(e) > -c(e)$  on edges  $e \in P^-$ . Here  $P = P^+ \cup P^-$ , where  $P^+$  consists of those  $e \in P$  whose orientations agree with the orientation of  $P$  from  $s$  to  $t$ . Let

$$(1.6) \quad \epsilon = \min[\min_{e \in P^+} (c(e) - x(e)), \min_{e \in P^-} (c(e) + x(e))] > 0$$

and define

$$(1.7) \quad x'(e) = \begin{cases} x(e), & e \notin P, \\ x(e) + \epsilon, & e \in P^+, \\ x(e) - \epsilon, & e \in P^-. \end{cases}$$

Then  $x'$  is a feasible flow from  $s$  to  $t$  of amount  $\alpha + \epsilon$ , contradicting the assumption that  $x$  was a maximum feasible flow. Hence  $t \in \bar{U}$ , as asserted. Let  $K$  be the set of edges joining  $U$  and  $\bar{U}$ , and write  $K = K^+ \cup K^-$ , where  $K^+(K^-)$  consists of those edges of  $K$  with positive (negative) end in  $U$ . Then  $K$  is a cut separating  $s$  and  $t$ , and it follows from the definition of  $U$  that  $x(e) = c(e)$  for  $e \in K^+$ ,

$x(e) = -c(e)$  for  $e \in K^-$ . Hence equality holds in (1.4).

Notice that the proof shows that a flow  $x$  is maximum if and only if there is no  $x$ -augmenting path from  $s$  to  $t$  (i.e., a path  $P$  such that (1.7) yields a feasible flow  $x'$ ).

If we assume that the capacity function  $c$  is integral- (or rational-) valued, the proof provides a good algorithm for constructing a maximum flow. We can begin the computation with any integral-valued feasible flow from  $s$  to  $t$ , e.g.,  $x(e) = 0$  all  $e \in E$ . We then institute a search for a flow-augmenting path using the prescription of (1.5a) and (1.5b). A good way to apply this prescription is to fan out from  $s$  to all its neighboring vertices that can be put into  $U$  using (1.5b); then repeat the process by selecting one of these vertices, scanning it for all its neighbors not yet in  $U$  that can now be put into  $U$ , and so on. This way of searching for a flow-augmenting path is called the "labeling process" in [8], where it is described in terms of assigning labels to vertices as we put them in  $U$ ; in terms of (1.5b), the label assigned to vertex  $v$  is  $u$ . (This simple process forms the basis for most of the network-programming algorithms described in [8].) If this search is successful in finding  $t$ , the flow increment  $\epsilon$  of (1.6) is a positive integer, and hence  $x'$  of (1.7) is again an integral-valued flow. If unsuccessful, the present flow is a maximum flow, and a minimum cut has been located. Thus the algorithm terminates, and at termination

we have constructed an integral maximum flow and a minimum cut.

Theorem 1.2. If the capacity function  $c$  is integral-valued, there is an integral maximum flow.

Theorem 1.2 is important in combinatorial applications of network flows.

While we have taken the capacity constraints (1.2) to be symmetric about the origin, there is no real need for this assumption. The constraints (1.2) can be changed to

$$(1.2') \quad b(e) \leq x(e) \leq c(e), \quad e \in E,$$

and handled in an analogous fashion provided they are feasible, that is, the constraint-set (1.1), (1.2') is nonempty. (Thus, for example, "one-way streets" can be incorporated in the model.) Even the feasibility question can be dealt with by an appropriate modification of the argument used in the proof of Theorem 1.1, or by applying a version of Theorem 1.1 to an enlarged network. For a detailed discussion of this and other extensions, e.g., capacities on vertices as well as edges, we refer to [8]. Here we shall simply state a typical feasibility theorem, the circulation theorem due to Hoffman [18].

Theorem 1.3. Let  $b(e) \leq c(e)$  for each edge  $e$  of a  
network  $G$  be given real numbers. The constraints (1.1')  
and (1.2') are feasible in  $G$  if and only if, for each  
subset  $U \subset V$ , we have

$$\sum_{e \in K^+} c(e) - \sum_{e \in K^-} b(e) \geq 0,$$

where  $K^+$  ( $K^-$ ) consists of those edges of  $G$  with positive  
(negative) end in  $U$  and negative (positive) end in  $V - U$ .

Minty [23] has distilled from the above proof of the max-flow min-cut theorem and from other network algorithms of Ford and Fulkerson [10, 11] a theorem about graphs, which Berge and Ghouila-Houri [1] have called "Lemme des Arcs Colorés." We call it the painting theorem. To state it, we require some definitions. A circuit  $C \subset E$  in graph  $G$  is a minimal closed path in  $G$ , that is, a set of edges which forms a closed path and is minimal with respect to this property. A cocircuit  $D \subset E$  is a minimal cut, that is, a set of edges whose deletion increases the number of connected components of  $G$  and is minimal with respect to this property. (In terms of the  $(0, \pm 1)$ -vertex-edge incidence matrix of an orientation of  $G$ , a circuit corresponds to a minimal dependent set of columns of the matrix, where "dependent" means "linearly dependent over the reals." If  $G$  is unoriented, and the vertex-edge matrix

is taken to be a  $(0, 1)$  - matrix, then a circuit corresponds to a minimal dependent set of columns, where "dependent" means "linearly dependent over the integers mod 2." A painting of  $G$  is a partition of the edges of  $G$  into three sets  $R$ ,  $W$ ,  $B$ , and the distinguishing of one edge of the set  $R$ . It may be viewed as painting the edges of  $G$  with three colors—red, white, blue—with one red edge being distinguished and painted dark red.

Theorem 1.4. Given a painting of an oriented graph  $G$ , precisely one of the following alternatives holds:

(i) There is a circuit in  $G$  containing the dark red edge but no white edge, in which all red edges are similarly oriented.

(ii) There is a cocircuit in  $G$  containing the dark red edge but no blue edge, in which all red edges are similarly oriented.

Proof: Let  $e' \sim (t, s)$  be the dark red edge. If  $e'$  is a loop, then (i) holds and (ii) fails, by the minimality of a cocircuit. If  $t \neq s$ , define a subset  $U \subset V$  recursively by the rules

(1.8a)  $s \in U$ ;

(1.8b) if  $u \in U$  and  $e \sim (u, v)$  is red or blue, then  $v \in U$ ; if  $u \in U$  and  $e \sim (v, u)$  is blue, then  $v \in U$ .

If  $t \in U$ , there is an elementary (minimal, simple) path from  $s$  to  $t$  of red and blue edges in which all red edges are oriented in the path direction. This path, together with edge  $e'$ , provides the circuit of (i). Conversely, if (i) holds, then  $t \in U$ . If  $t \notin U$ , consider the set of edges joining  $U$  to  $\bar{U} = V - U$ . These edges are either white or red, and any red edge is oriented from  $\bar{U}$  to  $U$ , as  $e'$  is. Delete these edges. The resulting graph has components  $U, \bar{U}_1, \dots, \bar{U}_k$  with  $t \in \bar{U}_1$ . The set of edges joining  $U$  and  $\bar{U}_1$  is the cocircuit of (ii). Conversely, if (ii) holds, then  $t$  cannot be in  $U$  via (1.8b).

To apply the painting theorem to the maximum flow problem, first add the return-flow edge  $e' \sim (t, s)$  to the network with  $c(e')$  large. Let  $x$  satisfy (1.1'), (1.2). Paint  $e'$  dark red. For other edges  $e$ : If  $c(e) = 0$ , paint  $e$  white; if  $x(e) = c(e) > 0$ , paint  $e$  red and reorient  $e$ ; if  $x(e) = -c(e) < 0$ , paint  $e$  red; if  $-c(e) < x(e) < c(e)$ , paint  $e$  blue. Alternative (i) of the painting theorem then leads to a flow-augmenting path, whereas (ii) leads to a minimum cut. In this application the white edges play a pale role—they could have been deleted once and for all. But there are other network-programming problems for which labeling algorithms that have been described [10, 11, 12, 23] can be viewed in terms of edge paintings; the role played by white edges is less passive in some of these.

Before leaving the discussion of maximum network flow, we mention an alternative formulation of the problem. This formulation is in terms of the path-edge incidence matrix of an unoriented graph; it was used in the first proof of the max-flow min-cut theorem [9]. Let  $\mathcal{P}$  be the collection of all paths from  $s$  to  $t$  in  $G$ . For each  $P \in \mathcal{P}$  and  $e \in E$  define an integer  $p(P, e) = 1$  or  $0$  according as  $e \in P$  or  $e \notin P$ . We call the resulting matrix the path-edge incidence matrix of  $G$ . Let  $y$  be a real-valued function with domain  $\mathcal{P}$  that satisfies

$$(1.9) \quad \sum_{P \in \mathcal{P}} y(P) p(P, e) \leq c(e), \quad e \in E,$$

$$(1.10) \quad y(P) \geq 0, \quad P \in \mathcal{P}.$$

Thus  $y(P)$  can be thought of as the magnitude of flow in  $P$ , and (1.9) says that the total amount of flow in  $e$  cannot exceed its capacity. Subject to (1.9), (1.10), we wish to maximize

$$(1.11) \quad \sum_{P \in \mathcal{P}} y(P).$$

This version of the problem might seem to be more restrictive, since if two paths  $P_1$  and  $P_2$  contain the same edge  $e$  in opposite directions, (1.9) insists that we add  $y(P_1)$  and  $y(P_2)$  instead of "cancelling flows in opposite directions."

The two formulations are equivalent, however.

If the capacity function  $c$  is integral valued, there is an integral-valued  $y$  satisfying (1.9), (1.10), and maximizing (1.11). An edge-form of Menger's theorem [20] can be deduced from this:

Theorem 1.5. Let  $G$  be an unoriented graph with two distinguished vertices  $s$  and  $t$ . The maximum number of edge-disjoint paths joining  $s$  and  $t$  is equal to the minimum number of edges in a cut separating  $s$  and  $t$ .

## 2. MINIMUM PATH

Let  $l(e)$  be a real nonnegative number associated with edge  $e$  of an unoriented, connected graph  $G$ . We shall think of  $l(e)$  as the length of edge  $e$ . The length of path  $P$  is

$$(2.1) \quad l(P) = \sum_{e \in P} l(e).$$

The second problem concerning two-terminal networks that we consider is the minimum path problem: to find a path joining  $s$  and  $t$  that has minimum length. There are several good methods known for doing this. We describe one below, but first we state and prove a theorem that is a path-cut dual of the max-flow min-cut theorem. Consider the maximum flow problem in terms of the path-edge incidence matrix. Suppose now that we form the cut-edge incidence matrix by defining  $d(K, e) = 1$  or  $0$  according as  $e \in K$  or  $e \notin K$ . Here  $K$  is a cut separating  $s$  and  $t$ . Let  $\mathcal{K}$  denote the

class of such cuts. Analogously to (1.9), (1.10), let  $y$  be a real-valued function with domain  $\mathcal{K}$  satisfying

$$(2.1) \quad \sum_{K \in \mathcal{K}} y(K) d(K, e) \leq \iota(e), \quad e \in E,$$

$$(2.2) \quad y(K) \geq 0, \quad K \in \mathcal{K}.$$

Again we wish to maximize

$$(2.3) \quad \sum_{K \in \mathcal{K}} y(K)$$

subject to these constraints.

The maximum value of (2.3) cannot exceed the length of a minimum path from  $s$  to  $t$ , because a path from  $s$  to  $t$  has some edge in common with each  $K \in \mathcal{K}$ .

Theorem 2.1. The maximum value of (2.3) subject to (2.1) and (2.2) is equal to the minimum path length from  $s$  to  $t$ .

The purely combinatorial version of (2.1) - (2.3) in which  $\iota(e) = 1$  all  $e \in E$  and  $y(K) = 0$  or  $1$  all  $K \in \mathcal{K}$ , asks for the maximum number of mutually disjoint cuts separating  $s$  and  $t$ . As was the case for the maximum flow problem, if  $\iota$  is integral valued, there is an integral-valued  $y$  that solves the linear program (2.1) - (2.3). This will follow from the proof given below. Hence the maximum number of disjoint cuts separating  $s$  and  $t$  is equal to the minimum number of edges in a path joining  $s$  and  $t$ .

Proof of Theorem 2.1. Let  $\pi(v)$  be the minimum path length from  $s$  to  $v$ , for all  $v \in V$ . Thus  $\pi(v) \geq 0$  and  $\pi(s) = 0$ . Let  $0 = \pi_0 < \pi_1 < \dots < \pi_n$  be the distinct values assumed by  $\pi$ . Partition  $V$  into  $n + 1$  parts  $V_0, V_1, \dots, V_n$ , where

$$V_i = \{v \in V \mid \pi(v) = \pi_i\}.$$

Thus  $s \in V_0$ . Suppose  $t \in V_k$ . We then single out  $k$  cuts  $K_1, K_2, \dots, K_k$  in  $\mathcal{K}$  by letting  $K_j$  be the set of edges joining vertices of  $\bigcup_{i=0}^{j-1} V_i$  and vertices of  $V - \bigcup_{i=0}^{j-1} V_i$ ,  $j = 1, 2, \dots, k$ . Define  $y(K_j) = \pi_j - \pi_{j-1}$ ,  $j = 1, 2, \dots, k$ , and  $y(K) = 0$  for other cuts  $K \in \mathcal{K}$ . Then  $y$  solves (2.1) - (2.3). To prove this, it suffices to show that  $y$  satisfies (2.1), since clearly  $y(K) \geq 0$  all  $K \in \mathcal{K}$ , and

$$\sum_{K \in \mathcal{K}} y(K) = \sum_{j=1}^k (\pi_j - \pi_{j-1}) = \pi_k - \pi_0 = \pi_k = \pi(t).$$

Thus consider an edge  $e$  joining a vertex  $u$  of  $V_i$  and a vertex  $v$  of  $V_j$ , where  $i < j \leq k$ , so that  $e$  belongs to each of the cuts  $K_{i+1}, \dots, K_j$ , but to no other cut having positive weight in  $y$ . Suppose that

$$y(K_{i+1}) + \dots + y(K_j) = \pi_j - \pi_i > t(e).$$

There is a path from  $s$  to  $u$  of length  $\pi_i$ ; adjoining  $e$  to

this path yields a path from  $s$  to  $v$  of length  $\pi_i + \ell(e) < \pi_i + (\pi_j - \pi_i) = \pi_j$ , a contradiction. If  $j > k$ , a similar contradiction results. Hence  $y$  satisfies (2.1) and solves (2.1) - (2.3).

For the case of a planar two-terminal network (that is, the graph  $G$  together with the additional edge  $e'$  joining the terminals  $s$  and  $t$  is a planar graph), where one can construct a dual two-terminal network in which source-sink paths correspond to cuts separating  $s$  and  $t$  in the primal network, the duality between the maximum flow problem and the minimum path problem was noted in [9], and was exploited in developing a max-flow algorithm for such networks, the "top-most path" method of [9]. Theorem 2.1 for arbitrary two-terminal networks is due to Robacker [27]. From the point of view of Part II of this paper, Theorem 2.1 and the max-flow min-cut theorem are abstractly the same.

We return now to the problem of constructing a minimum path joining  $s$  and  $t$ . The procedure we sketch here is a special case of a more general algorithm for constructing minimum cost flows in networks [11]. It evaluates the minimum path length  $\pi(v)$  from  $s$  to  $v$  for all  $v \in V$ , and hence provides a solution  $y$  to (2.1) - (2.3). We may suppose in the description that there are no loops or multiple edges in  $G$ . If edge  $e$  has ends  $u, v$ , we write the unordered pair  $(u, v)$  for  $e$  and  $\ell(u, v)$  for  $\ell(e)$ .

To start out, take  $\pi(s) = 0$ . Next look at all edges  $(s, v)$  and find the minimum value of  $\ell(s, v)$  for such edges. If  $v$  is a vertex yielding this minimum, set  $\pi(v) = \ell(s, v)$ .

The general step of the computation is as follows. Suppose that  $\pi(u)$  has been defined for  $u \in U \subset V$ . Let  $\bar{U} = V - U$  and compute

$$(2.4) \quad \min_{u \in U, v \in \bar{U}} [\pi(u) + \iota(u, v)] = \delta.$$

If the minimum in (2.4) is achieved for an edge  $(u, v)$ , set  $\pi(v) = \delta$ . Repeat the general step until  $\pi(v)$  has been defined for all  $v \in V$ . The number  $\pi(v)$  defined in this way is the minimum path length from  $s$  to  $v$ . A convenient way to do the calculation is to assign to vertex  $v$  the label  $(u, \pi(v))$ , where  $u$  is some vertex for which the minimum in (2.4) is achieved. A minimum path from  $s$  to  $v$  can then be found by backtracking from  $v$  to  $s$  as directed by first members of the labels.

At the conclusion of the computation, the numbers  $\pi(v)$  satisfy the inequalities

$$(2.5) \quad -\iota(u, v) \leq \pi(v) - \pi(u) \leq \iota(u, v)$$

for all edges  $(u, v)$  of  $G$ , and maximize  $\pi(t) - \pi(s)$  subject to (2.5). If we interpret  $\iota(u, v)$  as the cost of transporting a unit of some commodity over edge  $(u, v)$ , the number  $\pi(v)$  can be given the economic interpretation of a price placed on a unit of the commodity at location  $v$ . Inequalities (2.5) then say that no profit can be made by purchasing a

unit of the commodity at  $u$  and transporting it to  $v$  or vice versa. Subject to these restrictions, the price difference  $\pi(t) - \pi(s)$  is to be maximized. Thus the maximum value of  $\pi(t) - \pi(s)$  subject to (2.5) is equal to the minimum path cost from  $s$  to  $t$ . In another interpretation, Duffin has called this result the "max-potential equals min-work" theorem [5].

The assumption that edge lengths are nonnegative has been used in an essential way in this section. If edge lengths are allowed to be negative, and if we ask for a minimum length simple path joining two vertices, the problem is much harder. There are no known good algorithms for constructing such a path.

### 3. MAXIMUM CAPACITY PATH

Again we consider a two-terminal unoriented network  $G$  with source  $s$ , sink  $t$ , and capacity function  $c$ . This time we wish to find a path  $P$  from  $s$  to  $t$  that has the largest flow capacity of all such paths, i.e., we want to find a  $P$  that achieves

$$(3.1) \quad \max_{P \in \mathcal{P}} \min_{e \in P} c(e),$$

where  $\mathcal{P}$  is the class of all paths joining  $s$  and  $t$ . We call this the maximum capacity path problem.

This bottleneck problem has been considered in [13, 19, 26]. It is related to the minimum path problem in the sense

that methods for solving the latter can be modified to solve it. But here we shall describe another easy way of solving the problem, one that extends to blocking systems (Part III). This method of solution might be termed the "threshold method." It leads to the following min-max theorem concerning paths and cuts [13].

Theorem 3.1. Let  $G$  be a network with capacity function  $c$  and terminals  $s$  and  $t$ . Then

$$(3.2) \quad \max_{P \in \mathcal{P}} \min_{e \in P} c(e) = \min_{K \in \mathcal{K}} \max_{e \in K} c(e),$$

where  $\mathcal{P}$  is the class of all paths joining  $s$  and  $t$  and  $\mathcal{K}$  is the class of all cuts separating  $s$  and  $t$ .

Proof. If  $P \in \mathcal{P}$  and  $K \in \mathcal{K}$ , then  $P \cap K$  is nonempty. Let  $e' \in P \cap K$ . Then

$$\min_{e \in P} c(e) \leq c(e') \leq \max_{e \in K} c(e).$$

It follows that

$$(3.3) \quad \max_{P \in \mathcal{P}} \min_{e \in P} c(e) \leq \min_{K \in \mathcal{K}} \max_{e \in K} c(e).$$

To establish equality in (3.3), we can proceed as follows. Let  $c_1 > c_2 > \dots > c_n$  be the distinct values assumed by the capacity function, and let  $c_0$  be large.

Let  $G_i$  be the network obtained from  $G$  by deleting all edges  $e$  satisfying  $c(e) < c_i$ ,  $i = 0, 1, \dots, n$ . Thus  $G_0$  has no edges, and  $G_n = G$ . Suppose  $G_j$  is the first  $G_i$  that contains a path joining  $s$  and  $t$ . (We are tacitly assuming that  $\mathcal{P}$  is nonempty, although an appropriate interpretation of (3.2) holds if this isn't so.) Since  $G_j$  has a path  $P \in \mathcal{P}$  and  $G_{j-1}$  contains no path in  $\mathcal{P}$ , we have  $\min_{e \in P} c(e) = c_j$ . On the other hand, the edges deleted from  $G$  in forming  $G_{j-1}$  contain a cut  $K \in \mathcal{K}$ , whereas the edges deleted from  $G$  in forming  $G_j$  contain no cut in  $\mathcal{K}$ , and thus  $\max_{e \in K} c(e) = c_j$ . Consequently equality holds in (3.3).

Thus to solve the maximum capacity path problem, we lower the threshold for edge capacities until a path joining  $s$  and  $t$  is produced. There are good algorithms for recognizing when this happens.

Notice that no use is made of the fact that  $c(e) \geq 0$ . Indeed the solution depends only on the ordering of the edge numbers  $c(e)$ , not on their magnitudes.

An appropriate version of the threshold method can be used to locate a flow-augmenting path that yields the largest flow increment (1.6). Thus one way to solve the maximum flow problem is to successively find maximum capacity flow-augmenting paths by a threshold method.

One can also show

$$(3.4) \quad \min_{P \in \mathcal{P}} \max_{e \in P} c(e) = \max_{K \in \mathcal{K}} \min_{e \in K} c(e).$$

For an interpretation, think of  $G$  as a highway map with  $c(e)$  being the maximum elevation encountered in driving over edge  $e$ .

#### 4. LENGTH-WIDTH INEQUALITY.

Duffin [5] has defined the notions of "extremal length" and "extremal width" for two-terminal networks having edge resistances and has shown that these are reciprocal quantities. From this relationship he deduced a certain inequality concerning paths and cuts for a two-terminal network in which each edge has associated with it two nonnegative numbers  $l(e)$  and  $w(e)$ , the length and width of  $e$ . An earlier, purely combinatorial version of this inequality in which  $l(e) = w(e) = 1$  is due to Moore and Shannon [25]. This version says that if  $\lambda$  is the least number of edges in a path joining  $s$  and  $t$  and  $\omega$  is the least number of edges in a cut separating  $s$  and  $t$ , then  $\lambda\omega$  is less than or equal to the number of edges in the graph. More generally, let

$$(4.1) \quad \lambda = \min_{P \in \mathcal{P}} l(P) = \min_{P \in \mathcal{P}} \sum_{e \in P} l(e),$$

$$(4.2) \quad \omega = \min_{K \in \mathcal{K}} w(K) = \min_{K \in \mathcal{K}} \sum_{e \in K} w(e),$$

where  $\mathcal{P}$  is the class of all paths joining  $s$  and  $t$ ,  $\mathcal{K}$  is the class of all cuts separating  $s$  and  $t$ . The number  $\lambda$

is called the length of  $G$ ,  $\omega$  the width of  $G$ , relative to  $s$  and  $t$ . The length-width inequality asserts that

$$(4.3) \quad \lambda \omega \leq \sum_{e \in E} \iota(e) w(e).$$

A proof of (4.3) can be given using either the max-flow min-cut theorem or its path-cut dual. We use the former approach. Interpret  $w(e)$  as the flow-capacity of  $e$ . Then by the max-flow min-cut equality, there is a flow from  $s$  to  $t$  of magnitude  $\omega$ . It follows that there is a function  $y$  defined on  $\mathcal{P}$  satisfying (1.9), (1.10), and

$$\sum_{P \in \mathcal{P}} y(P) = \omega.$$

Thus

$$\begin{aligned} \lambda \omega &= \lambda \sum_{P \in \mathcal{P}} y(P) \leq \sum_{P \in \mathcal{P}} \iota(P) y(P) = \sum_{P \in \mathcal{P}} \sum_{e \in P} \iota(e) y(P) \\ &\leq \sum_{e \in E} \iota(e) \sum_{P \in \mathcal{P}} y(P) p(P, e) \leq \sum_{e \in E} \iota(e) w(e). \end{aligned}$$

Although the length-width inequality appears weak, we shall point out in Part III that the existence of a length-width inequality for a blocking system implies the max-flow min-cut equality for the system.

## PART II - FRAMES

Our aim in this part of the paper is to indicate how the theorems of Part I can be generalized to frames of subspaces of Euclidean  $n$ -space. (We shall define a frame later on. But it should be mentioned here that the word "frame" was used by Tutte in some of his early work on chain-groups and matroids in place of the word "matroid". We appropriate it, with his permission, for a more restrictive use.) The notion of a frame is closely related to that of a matric matroid. Indeed a frame can be viewed as the structure obtained just prior to the matroid in making the transition from matrix to its matroid.

Matroids were introduced by Whitney [35] as a generalization of dependence properties in graphs or in matrices. There is now an extensive and deep theory of matroids, mostly due to Tutte [30, 31, 32, 33, 34]. We require only the more elementary parts of this theory. (Certainly Tutte's Introduction to the Theory of Matroids [34] would suffice.)

The generalization from Part I to Part II can be described roughly as that obtained by replacing the vertex-edge incidence matrix of an oriented graph by an arbitrary real matrix. (More generally, we could consider matrices over any ordered field.) Thus we pass from the special network programs of Part I to general linear programs.

Associated with every linear program there is a dual program. Associated with every matroid there is a dual

matroid. Associated with every frame there is a dual frame. Frame duality provides a bridge between matroid duality and linear programming duality. The basic concept underlying duality in all three instances is orthogonality.

Although the material of this part of the paper was developed independently by the writer, we doubt that much of it is new. A recent paper by Rockafellar [28] contains a similar development, for example. Our attention has also been called to work of Camion [2], and to a forthcoming book on networks by Iri. Most of the notions and some of the results are either explicit or implicit in Tutte's work on matroids. We believe that our treatment of the generalized maximum flow problem and the resulting length-width inequality for real matrices may be new, however.

### 1. FRAMES OF REAL SUBSPACES

Let  $\mathcal{R}$  be an arbitrary subspace of  $n$ -dimensional Euclidean space  $\mathcal{R}^n$ . For the correspondence with Part I, a vector  $X = (x_1, x_2, \dots, x_n)$  in  $\mathcal{R}^n$  should be thought of as a real-valued function on a finite set of "edges"  $E = \{e_1, e_2, \dots, e_n\}$  that maps  $e_i$  into  $x_i$ , and  $\mathcal{R}$  should be viewed as the row space of an  $m$  by  $n$  real matrix  $A = (a_{ij})$ , the "generalized vertex-edge incidence matrix".

Let  $Y = (y_1, y_2, \dots, y_n)$  be a vector of  $\mathcal{R}$ . The support  $S(Y)$  of  $Y$  consists of those  $e_i \in E$  such that  $y_i \neq 0$ . A vector  $Y \in \mathcal{R}$  is called an elementary vector of  $\mathcal{R}$  if it

is nonzero and if there is no nonzero vector  $X \in \mathcal{R}$  such that  $S(X)$  is a proper subset of  $S(Y)$ . Thus if  $X$  and  $Y$  are two elementary vectors of  $\mathcal{R}$  having the same support, then  $X$  is a nonzero multiple of  $Y$ . Consequently we may associate with  $\mathcal{R}$  a unique, finite set of lines, each line being determined by an elementary vector of  $\mathcal{R}$ . We call this collection of lines the frame  $\mathcal{F} = \mathcal{F}(\mathcal{R})$  of  $\mathcal{R}$ , and sometimes refer to an elementary vector  $F$  of  $\mathcal{R}$  as a frame-vector of  $\mathcal{R}$ .

Let  $X$  and  $Y$  be vectors of  $\mathcal{R}$ . The vector  $X$  conforms to  $Y$  if  $x_i y_i > 0$  whenever  $x_i \neq 0$ . In particular,  $S(X) \subset S(Y)$ .

Lemma 1.1. Let  $Y$  be a nonzero vector of  $\mathcal{R}$ . There exists an elementary vector  $F$  of  $\mathcal{R}$  that conforms to  $Y$ .

Proof: If not, select  $Y = (y_1, y_2, \dots, y_n) \in \mathcal{R}$  so that no elementary vector of  $\mathcal{R}$  conforms to  $Y$ , and so that the number of elements in  $S(Y)$  is as small as possible consistent with this condition. Let  $X = (x_1, x_2, \dots, x_n)$  be an elementary vector of  $\mathcal{R}$  such that  $S(X) \subset S(Y)$ . Let  $I \subset E$  denote the set of  $e_i \in E$  such that  $y_i$  and  $x_i$  have opposite signs. Thus  $I$  is nonempty. Consider the vector  $Z = Y + aX$ , where

$$a = \min_{e_i \in I} \left( -\frac{y_i}{x_i} \right) > 0.$$

The vector  $Z$  conforms to  $Y$  and  $S(Z)$  is properly included in  $S(Y)$ . By the selection of  $Y$ , there is an elementary

vector  $F$  conforming to  $Z$ . But then  $F$  conforms to  $Y$ . This contradiction establishes the lemma.

An important consequence of Lemma 1.1 is that any non-zero vector  $Y \in \mathcal{R}$  can be written as a sum

$$(1.1) \quad Y = F_1 + F_2 + \dots + F_k$$

of elementary vectors of  $\mathcal{R}$ , where each elementary vector  $F_i$  in (1.1) conforms to  $Y$ , and two elementary vectors  $F_i, F_j$  with  $i \neq j$  lie on distinct frame-lines of  $\mathcal{R}$ . We call (1.1) a conformal frame decomposition of  $Y$ . In general, such a decomposition is far from unique, of course.

We return now to the matrix  $A = (a_{ij})$  whose rows generate  $\mathcal{R}$ . A (column) pivot on an element  $a_{kl} \neq 0$  of  $A$  is a sequence of elementary row operations on  $A$  that transforms  $A$  into a matrix  $A' = (a'_{ij})$  in which  $a'_{kl} = 1$ ,  $a'_{il} = 0$  for  $i \neq k$ . Starting with  $A$ , we can produce from it by a sequence of column pivots and deletions of zero rows a matrix  $R$  whose columns can be permuted to have the form

$$(1.2) \quad (I, B).$$

If  $A$  has rank  $r$ , then  $R$  is  $r$  by  $n$ , the rows of  $R$  are a basis for  $\mathcal{R}$ , and  $R$  contains an  $r$  by  $r$  permutation sub-matrix whose columns correspond to some  $S \subset E$ . Following

Tutte, we refer to such a matrix  $R$  as a standard representative matrix of  $\mathcal{K}$ . Note that each row of  $R$  is an elementary vector of  $\mathcal{K}$ . The following theorem asserts that, conversely, any elementary vector of  $\mathcal{K}$  can be obtained from  $A$  by a finite sequence of pivots.

Theorem 1.2. Let  $F$  be an elementary vector of  $\mathcal{K}$ . Then there exists a standard representative matrix  $R$  of  $\mathcal{K}$  having a multiple of  $F$  as one of its rows.

Proof. Extend  $F$  to a basis  $\mathcal{B}$  of  $\mathcal{K}$ , and write the resulting collection of vectors as a matrix having  $F$  as its first row, say. Pivot on a nonzero coordinate of  $F$ . Consider the second row of the transformed matrix. This row has a nonzero coordinate in one of the columns corresponding to zero coordinates of the first row, for otherwise either  $F$  would not be elementary or  $\mathcal{B}$  would not be a basis. Pivot on such an element. Repetition of this process produces a standard representative matrix  $R$  of  $\mathcal{K}$  having a multiple of  $F$  as its first row.

In particular, an elementary vector of  $\mathcal{K}$  can have at most  $n - r + 1$  nonzero coordinates.

Notice also that if  $\mathcal{K}$  and  $\mathcal{L}$  are subspaces having the same frame  $\mathcal{B}(\mathcal{K}) = \mathcal{B}(\mathcal{L})$ , then  $\mathcal{K} = \mathcal{L}$ .

## 2. MATROIDS

A matroid is a purely combinatorial structure defined on a finite set  $E$ . There are a number of equivalent axiom

systems for matroids. One in terms of "circuits" is as follows. Let  $\mathcal{C}$  be a finite family of nonempty subsets of  $E$ . Members of  $\mathcal{C}$  are the circuits of a matroid  $(E, \mathcal{C})$  if the following axioms hold:

(2.1) No member of  $\mathcal{C}$  is a proper subset of another.

(2.2) Let  $e_1$  and  $e_2$  be distinct members of  $E$ , and suppose  $C_1$  and  $C_2$  are members of  $\mathcal{C}$  such that  $e_1 \in C_1 \cap C_2$  and  $e_2 \in C_1 - C_2$ . Then there exists  $C_3 \in \mathcal{C}$  such that  $e_2 \in C_3 \subset (C_1 \cup C_2) - \{e_1\}$ .

The motivation comes from graphs. Let  $E$  be the set of edges of an unoriented graph  $G$ . Then the collection  $\mathcal{C}$  of (graph) circuits of  $G$  satisfies (2.1), (2.2), and thus  $(E, \mathcal{C})$  is a matroid. Such a matroid is graphic. The collection  $\mathcal{D}$  of cocircuits of  $G$  also satisfies (2.1), (2.2), and thus forms a matroid  $(E, \mathcal{D})$ . Such a matroid is cographic. For another important example, consider the row space  $\mathcal{R}$  of the  $m$  by  $n$  matrix  $A$ . Take  $E = \{e_1, e_2, \dots, e_n\}$ . Then the collection  $\mathcal{C}$  of supports of frame-vectors of  $\mathcal{R}$  satisfies (2.1), (2.2) and is consequently a matroid  $(E, \mathcal{C})$ . Such a matroid is called a real matric matroid.

Associated with every matroid  $(E, \mathcal{C})$  there is a unique dual matroid  $(E, \mathcal{C}^*)$ . A subset of  $E$  is a member of  $\mathcal{C}^*$

if and only if the cardinality of its intersection with every element of  $\mathcal{C}$  is not equal to 1, and it is minimal with respect to this property. The dual of the dual is the primal:  $(E, \mathcal{C}^{**}) = (E, \mathcal{C})$ . In case  $(E, \mathcal{C})$  is a graphic matroid, the cographic matroid  $(E, \mathcal{D})$  is the dual:  $(E, \mathcal{D}) = (E, \mathcal{C}^*)$ ,  $(E, \mathcal{D}^*) = (E, \mathcal{C})$ . If  $(E, \mathcal{C})$  is a real matrix matroid arising from a subspace  $\mathcal{R}$ , the dual matroid is the real matrix matroid obtained from the orthogonal complement  $\mathcal{R}^*$  of  $\mathcal{R}$ . Thus if  $\mathcal{F}$  is the frame of  $\mathcal{R}$ , we call the frame  $\mathcal{F}^*$  of  $\mathcal{R}^*$  the dual of  $\mathcal{F}$ . If  $\mathcal{R}$  has standard representative matrix  $R = (I_r, B)$ , then a standard representative matrix for  $\mathcal{R}^*$  is  $R^* = (B^T, -I_{n-r})$ . A frame-vector of  $\mathcal{R}$  can be viewed as representing the coefficients of a minimal linear dependency among columns of  $R^*$ .

Let  $A = (a_{ij})$  be the vertex-edge incidence matrix of an oriented graph  $G$ . It is well-known that the matrix  $A$  has the total unimodularity property: every square submatrix of  $A$  has determinant 0, 1, or -1. One can deduce from this that each elementary vector of the row space  $\mathcal{R}$  of  $A$  is a multiple of a vector having coordinates 0, 1, or -1. Such a vector is called primitive. Conversely, if a subspace  $\mathcal{R}$  has the property that each elementary vector of  $\mathcal{R}$  is a multiple of a primitive vector, then  $\mathcal{R}$  is the row space of some totally unimodular matrix  $A = (a_{ij})$ . In particular,  $a_{ij} = 0, 1, \text{ or } -1$ . Such a space  $\mathcal{R}$  is called

regular and the corresponding matroid is a regular matroid.

Thus regular matroids are precisely those real matric matroids generated by totally unimodular matrices. The dual of a regular matroid is regular. A dual pair of regular matroids is called a "digraphoid" in [24].

(It should be remarked, though we make no use of it here, that Tutte has shown that a regular matroid is a binary matric matroid, that is, a matroid generated by a matrix over the field of two elements, and has characterized regular matroids as a subset of the binary matric matroids. This characterization, which is in terms of certain excluded matroid minors—a matroid minor is not the same thing as a matrix minor—is deeper than the one above, also due to Tutte, of regular matroids as a subset of real matric matroids. It can also be shown, as was pointed out to the writer by Edmonds, that a matroid is regular if and only if it is both a real matric matroid and a binary matric matroid. From this one can deduce that a  $(0, \pm 1)$ -matrix  $(I, B)$  is totally unimodular if and only if the binary rank of any subset  $S$  of its columns is equal to the real rank of  $S$ . This can also be proved directly. It is also possible to give a characterization of regular matroids among those real matric matroids generated by  $(0, \pm 1)$ -matrices in terms of a single excluded matroid minor: namely, exclude the self-dual matroid on a set of four elements, every triple of which is a circuit. The problem of characterizing

regular matroids among all real matric matroids in terms of excluded matroid minors appears to be open, as does the more fundamental problem of giving necessary and sufficient conditions in order that two real matrices generate the same matroid.)

The real matric matroid generated by the vertex-edge incidence matrix  $A$  of an oriented graph is a regular matroid. The nonzero coordinates of an elementary vector  $F$  of the row space  $\mathcal{R}$  of  $A$  pick out a cocircuit in the graph, two edges being similarly oriented in this cocircuit if the corresponding coordinates of  $F$  have the same sign. Conversely, each cocircuit of the graph can be exhibited in this way as an elementary vector of  $\mathcal{R}$ . On the other hand, nonzero coordinates of an elementary vector of  $\mathcal{R}^*$  pick out a circuit in the graph, two edges being similarly oriented in this circuit if the corresponding coordinates have the same sign, and each circuit of the graph can be exhibited in this way.

### 3. GENERALIZED FLOWS AND CUTS

Let  $A = (a_{ij})$  be an  $m$  by  $n$  real matrix having row space  $\mathcal{R}$ . For each  $e_j \in E = \{e_1, e_2, \dots, e_n\}$ , let  $c_j$  be a nonnegative real number, the capacity of  $e_j$ . In analogy with (1.1') and (1.2) of Part I, we define a (feasible) flow  $X$  on  $A$  to be a vector  $X = (x_1, x_2, \dots, x_n)$  that satisfies the linear homogeneous equations

$$(3.1) \quad \sum_{j=1}^n a_{ij} x_j = 0, \quad i = 1, 2, \dots, m,$$

and inequalities

$$(3.2) \quad -c_j \leq x_j \leq c_j, \quad j = 1, 2, \dots, n.$$

Thus  $X \in \mathcal{R}^*$ . Clearly feasible flows exist, e.g.  $X = 0$ .

The analogue of the maximum flow problem is to find a feasible flow  $X$  on  $A$  that maximizes some specified component of  $X$ , say  $x_1$ , where  $c_1 = \infty$ . We call such a flow a maximum  $e_1$ -flow.

Let  $K = (-1, k_2, \dots, k_n)$  be an elementary vector of  $\mathcal{R}$ . (Such elementary vectors exist unless the first column of  $A$  consists entirely of 0's—this corresponds to the graphic case in which  $e_1$  is a loop.) We say that  $K$  is an  $e_1$ -cut. There are finitely many such. The capacity of an  $e_1$ -cut  $K$  is defined to be

$$(3.3) \quad \sum_{\substack{j=2 \\ k_j > 0}}^n k_j c_j - \sum_{\substack{j=2 \\ k_j < 0}}^n k_j c_j = \sum_{j=2}^n |k_j c_j|.$$

If  $X$  is a feasible flow and  $K$  an  $e_1$ -cut, then, since  $X \in \mathcal{R}^*$  and  $K \in \mathcal{R}$ , we have

$$\sum_{j=1}^n x_j k_j = 0,$$

and hence, by (3.2),

$$(3.4) \quad x_1 = \sum_{j=2}^n x_j k_j \leq \sum_{j=2}^n |k_j c_j|.$$

Theorem 3.1. The maximum value of  $x_1$  subject to (3.1) and (3.2) is equal to the minimum capacity of all  $e_1$ -cuts.

Proof. It suffices to show that there is a flow and an  $e_1$ -cut for which equality holds in (3.4). A proof of this can be given using either the linear programming duality theorem [3, 16] or Dantzig's simplex method for solving linear programs [3]. We sketch the former approach. Let  $X = (x_1, x_2, \dots, x_n)$  be a maximum  $e_1$ -flow. The duality theorem for the linear program at hand then implies that there exists an  $m$ -vector  $(\pi_1, \pi_2, \dots, \pi_m)$  such that the following "optimality" properties hold:

$$(3.5) \quad 1 + \sum_{i=1}^m \pi_i a_{i1} = 0,$$

and, for  $j = 2, \dots, n$ ,

$$(3.6) \quad \sum_{i=1}^m \pi_i a_{ij} > 0 \Rightarrow x_j = c_j,$$

$$\sum_{i=1}^m \pi_i a_{ij} < 0 \Rightarrow x_j = -c_j.$$

Let

$$Y = \left( \sum_{i=1}^m \pi_i a_{i1}, \dots, \sum_{i=1}^m \pi_i a_{in} \right) = (-1, y_2, \dots, y_n).$$

Thus  $Y \in \mathcal{R}$ . By Lemma 1.1, there exists an elementary vector  $K = (-1, k_2, \dots, k_n)$  of  $\mathcal{R}$  that conforms to  $Y$ . The properties (3.6) then hold for  $K$  and imply that equality holds in (3.4). This proves Theorem 3.1.

The simplex method constructs a maximum  $e_1$ -flow and a minimum  $e_1$ -cut simultaneously. Indeed, the method proceeds by a sequence of pivots on  $A$ , and at termination yields a standard representative matrix  $R$  of  $\mathcal{R}$ , one of whose rows is an  $e_1$ -cut of minimum capacity.

If  $A$  is totally unimodular, then the coordinates of  $K$  in Theorem 3.1 are 0, 1, or -1, and we have a more purely combinatorial result: namely, the generalization of the max-flow min-cut theorem to regular matroids or digraphoids noted in [24]. Observe that the analogue of the integrity theorem, Theorem 1.2 of Part I, is valid for this case.

Just as for the case of flows in networks, the assumption of symmetric capacity constraints can easily be dispensed with in Theorem 3.1. The capacity constraints can be changed to  $b_j \leq x_j \leq c_j$ , and treated in a similar fashion, provided they are feasible. The capacity of an  $e_1$ -cut  $K = (-1, k_2, \dots, k_n)$  is then defined to be

$$(3.3') \quad \sum_{\substack{j=2 \\ k_j > 0}}^n k_j c_j + \sum_{\substack{j=2 \\ k_j < 0}}^n k_j b_j.$$

The feasibility question is most conveniently disposed of by the following "generalized circulation theorem," the analogue of Theorem 1.3, Part I.

Theorem 3.2. Let  $A = (a_{ij})$  be an  $m$  by  $n$  real matrix,  
and let  $b_j \leq c_j$ ,  $j = 1, 2, \dots, n$ , be given real numbers.  
The constraints

$$(3.7) \quad \sum_{j=1}^n a_{ij} x_j = 0, \quad i = 1, 2, \dots, m,$$

$$(3.8) \quad b_j \leq x_j \leq c_j, \quad j = 1, 2, \dots, n,$$

are feasible if and only if, for each elementary vector  
 $K = (k_1, k_2, \dots, k_n)$  in the row space  $\mathcal{R}$  of  $A$ , we have

$$(3.9) \quad \sum_{k_j > 0} k_j c_j + \sum_{k_j < 0} k_j b_j \geq 0.$$

Notice that (3.9) is really a finite set of inequalities, since we need only choose from each frame-line of  $\mathcal{R}$  one elementary vector and its negative in checking (3.9).

We turn next to the painting theorem for a real  $m$  by  $n$  matrix  $A = (a_{ij})$ . Here we paint the edges of  $E = \{e_1, e_2, \dots, e_n\}$ , i.e. the columns of  $A$ , with three colors—red,

white, blue—with one red edge being distinguished and painted dark red. Two edges  $e_i, e_j$  are similarly oriented in an elementary vector  $X = (x_1, x_2, \dots, x_n)$  of a subspace  $\mathcal{R} \subset \mathcal{R}^n$  if  $x_i x_j > 0$ ;  $X$  contains  $e_i$  if  $e_i \in S(X)$ .

Theorem 3.3. Given a painting of  $E = \{e_1, e_2, \dots, e_n\}$  and a real  $m$  by  $n$  matrix  $A = (a_{ij})$  having row space  $\mathcal{R}$ , precisely one of the following alternatives holds:

(i) There is an elementary vector  $X$  of  $\mathcal{R}^*$  containing the dark red edge but no white edge, in which all red edges are similarly oriented.

(ii) There is an elementary vector  $Y$  of  $\mathcal{R}$  containing the dark red edge but no blue edge, in which all red edges are similarly oriented.

Proof. Clearly both alternatives can't hold, since  $\mathcal{R}$  and  $\mathcal{R}^*$  are orthogonal.

Delete all white columns of  $A$ . Pivot on blue columns, one after another in any order, until no more such pivots are possible. (These operations correspond to the "deletions" and "contractions" of edges in graph or matroid theory.) Now delete all rows and columns of the resulting matrix that contain pivotal elements. Call the remaining matrix  $\bar{A}$ . Note that any blue columns of  $\bar{A}$  consist entirely of 0's. ( $\bar{A}$  generates a matroid minor of the matroid generated by  $A$ .) Let  $\bar{\mathcal{R}}$  denote the row space of  $\bar{A}$ . It follows from standard theorems on linear inequalities that (just) one of a pair of complementary orthogonal subspaces contains a nonnegative

vector whose first coordinate, say, is positive. Thus either  $\bar{\mathcal{R}}$  or  $\bar{\mathcal{R}}^*$  (but not both) contains a nonnegative vector whose dark red coordinate is positive. Suppose  $\bar{Z} \in \bar{\mathcal{R}}^*$  is such a vector. Then  $\bar{Z}$  can be extended to a vector  $Z \in \mathcal{R}^*$  such that white coordinates of  $Z$  are all zero. In this case (i) holds, by Lemma 1.1. Suppose that  $\bar{Z} \in \bar{\mathcal{R}}$  is a nonnegative vector whose dark red coordinate is positive. In this case  $\bar{Z}$  can be extended to a vector  $Z \in \mathcal{R}$  such that all blue coordinates of  $Z$  are zero. In this case (ii) holds, by Lemma 1.1.

If  $A$  is totally unimodular, the elementary vectors  $X$  and  $Y$  of Theorem 3.3 can be taken to be primitive, and Theorem 3.3 reduces to the painting theorem for digraphoids [24].

We return now to the version of Theorem 3.1 with capacity constraints  $b_j \leq x_j \leq c_j$ . How general is the class of linear programs encompassed by this theorem? The answer is not hard to see: it includes all linear programs. For, as is well known, any linear program can be put in the form

$$(3.10) \quad \sum_{j=1}^n a_{ij}x_j = b_i, \quad i = 1, 2, \dots, m,$$

$$x_j \geq 0, \quad j = 1, 2, \dots, n,$$

$$\text{maximize} \quad \sum_{j=1}^n c_j x_j.$$

Introducing new variables  $x_0$  and  $x_{n+1}$ , we see that (3.10) is equivalent to

$$(3.11) \quad \sum_{j=1}^n a_{ij}x_j - b_i x_{n+1} = 0,$$

$$x_0 - \sum_{j=1}^n c_j x_j = 0,$$

$$-\infty \leq x_0 \leq \infty,$$

$$0 \leq x_j \leq \infty, \quad j = 1, 2, \dots, n,$$

$$1 \leq x_{n+1} \leq 1,$$

maximize  $x_0$ .

The program (3.11) is a maximum flow problem on a subspace of  $\mathcal{R}^{n+2}$ .

Still following the discussion of Part I, Section 1, let us look now at the general version of the path-edge formulation of the maximum flow problem. Is there an analogue of (1.9), (1.10), and (1.11) for an arbitrary real matrix? We shall see that there is. Consider the matrix whose rows consist of all elementary vectors of  $\mathcal{R}^*$  of the form  $(1, p_2, \dots, p_n)$ . Let  $(p_{kj})$ ,  $k = 1, 2, \dots, s$ ,  $j = 1, 2, \dots, n$ , denote this matrix and let  $(|p_{kj}|)$  be the matrix obtained by taking absolute values of elements. We want

to show that the programs

$$(3.12) \quad \sum_{k=1}^s y_k \cdot |p_{kj}| \leq c_j, \quad j = 1, 2, \dots, n,$$

$$y_k \geq 0,$$

$$\text{maximize} \quad \sum_{k=1}^s y_k \cdot |p_{k1}| = \sum_{k=1}^s y_k,$$

and

$$(3.13) \quad \sum_{j=1}^n a_{ij} x_j = 0, \quad i = 1, 2, \dots, m,$$

$$-c_j \leq x_j \leq c_j, \quad j = 1, 2, \dots, n,$$

$$\text{maximize} \quad x_1,$$

are equivalent. Here we take  $c_1 = \infty$ . Given a feasible solution  $Y = (y_1, y_2, \dots, y_s)$  of (3.12), define

$$x_j = \sum_{k=1}^s y_k p_{kj}. \quad \text{Then } -c_j \leq x_j \leq c_j \text{ and}$$

$$\sum_{j=1}^n a_{ij} x_j = \sum_{k=1}^s \left( \sum_{j=1}^n a_{ij} p_{kj} \right) y_k = 0.$$

Conversely, given a feasible solution  $X = (x_1, x_2, \dots, x_n)$  of (3.13), we use the conformal frame decomposition (1.1) to write

$$(3.14) \quad X = y_1 P_1 + \dots + y_l P_l + F_1 + \dots + F_h, \quad y_k > 0,$$

where  $P_1, \dots, P_l$  are the first  $l$  rows, say, of the matrix  $(p_{kj})$ , and each elementary vector in (3.14) conforms to  $X$ . Define  $y_k = 0$  for the remaining rows of  $(p_{kj})$ . It follows that

$$\sum_{k=1}^s y_k \cdot |p_{kj}| \leq c_j.$$

Thus (3.12) and (3.13) are equivalent programs.

In particular, if  $A$  is totally unimodular, then  $(|p_{kj}|)$  is a  $(0, 1)$ -incidence matrix, and an integral  $X$  in (3.14) yields an integral  $Y$  solving (3.12). Thus integral capacities lead to integral solutions in both programs. This observation establishes an analogue of Theorem 1.5, Part I. That is, an analogue of the edge form of Menger's theorem is valid for regular matroids. This has previously been shown by Minty in [24].

It seems likely that the relationship between (3.12) and (3.13) has implications for what is called the "decomposition principle" in linear programming. We shall not pursue this point here.

The only other problem from Part I that we want to examine in the context of Part II is the length-width inequality. (The generalized minimum path problem is the frame-dual of the generalized maximum flow problem and thus presents nothing new. Part III will be devoted to a

very general combinatorial analogue of the maximum capacity path problem.) Let  $A = (a_{ij})$  be a real  $m$  by  $n$  matrix with row space  $\mathcal{R}$ , and suppose  $\iota_j, w_j$  are given nonnegative numbers for  $j = 2, \dots, n$ . Consider the collection  $\mathcal{P} = \{P_1, \dots, P_r\}$  of all elementary vectors of  $\mathcal{R}^*$  that have first coordinate 1, and the collection  $\mathcal{K} = \{K_1, \dots, K_s\}$  of all elementary vectors of  $\mathcal{R}$  that have first coordinate 1. Let

$$(3.15) \quad \lambda = \min_{1 \leq i \leq r} \sum_{j=2}^n |p_{ij} \iota_j|,$$

$$(3.16) \quad \omega = \min_{1 \leq k \leq s} \sum_{j=2}^n |k_{hj} w_j|,$$

where

$$P_i = (1, p_{i2}, \dots, p_{in}), \quad i = 1, 2, \dots, r,$$

$$K_h = (1, k_{h2}, \dots, k_{hn}), \quad h = 1, 2, \dots, s.$$

We call  $\lambda$  the  $e_1$ -length of  $A$ , and call  $\omega$  the  $e_1$ -width of  $A$ .

Theorem 3.4. Let  $A = (a_{ij})$  be an  $m$  by  $n$  real matrix having  $e_1$ -length  $\lambda$  relative to  $\iota_j \geq 0, j = 2, \dots, n$ , and  $e_1$ -width  $\omega$  relative to  $w_j \geq 0, j = 2, \dots, n$ . Then

$$(3.17) \quad \lambda \omega \leq \sum_{j=2}^n \iota_j w_j.$$

Proof. By Theorem 3.1, Part II (the generalized max-flow min-cut theorem) and the equivalence of (3.12) and (3.13), it follows from (3.16) that there exists a nonnegative  $r$ -vector  $Y = (y_1, y_2, \dots, y_r)$  such that

$$(3.17) \quad \sum_{i=1}^r |y_i p_{ij}| \leq w_j, \quad j = 2, \dots, n,$$

$$\sum_{i=1}^r y_i = w.$$

Thus we have

$$\lambda w = \lambda \sum_{i=1}^r y_i \leq \sum_{i=1}^r \sum_{j=2}^n |p_{ij} t_j y_i| \leq \sum_{j=2}^n w_j t_j$$

by (3.18), (3.15), and (3.17), respectively.

Again if  $A$  is totally unimodular, then each  $P \in \mathcal{P}$  and  $K \in \mathcal{K}$  is primitive; taking  $t_j = w_j = 1$  gives a direct generalization of the Moore-Shannon theorem for graphs to totally unimodular matrices. In matroidal terms:

Corollary 3.5. Let  $(E, \mathcal{C})$  be a regular matroid on  $n$  edges. Let  $\lambda(e) + 1$  be the least number of edges in any circuit containing edge  $e$ ,  $w(e) + 1$  the least number of edges in any cocircuit (circuit of the dual matroid) containing  $e$ . Then  $\lambda(e)w(e) \leq n - 1$ .

### PART III - BLOCKING SYSTEMS

In Part I we described the maximum capacity path problem for a two-terminal network, gave a good algorithm for solving it, and presented a min-max theorem concerning paths and cuts for the problem. Similarly, Gross [17] has described a good algorithm and a min-max theorem for the "bottleneck assignment problem": Given a square array of real numbers, find a circling of entries with exactly one circle in each row and in each column so as to maximize the value of the smallest circled entry. For an interpretation, think of rows of the array as corresponding to men, columns to jobs on a serial assembly line, with the entry in row  $i$  and column  $j$  being the rate at which man  $i$  can process items if he is assigned to job  $j$ . The theorem established in [17] for this problem is the following: Let  $I = \{1, 2, \dots, n\}$ , let  $\mathcal{P}$  be the set of permutations of  $I$ , let  $|C|$  denote cardinality of  $C$ , and let  $a_{ij}$ ,  $i \in I$ ,  $j \in I$ , be real numbers. Then

$$\max_{P \in \mathcal{P}} \min_{i \in I} a_{i, P(i)} = \min_{\substack{A, B \subseteq I \\ |A| + |B| = n+1}} \max_{\substack{i \in A \\ j \in B}} a_{ij}.$$

The resemblance between these two min-max theorems is more than superficial. They are, in fact, special cases of a general theorem for a combinatorial structure which might be called a blocking system. These systems have arisen

in numerous contexts (see [21, 22, 29], for example), but the particular axiomatization and general min-max theorem presented in [7] and surveyed here, have apparently not been noted before.

### 1. AXIOMS AND EXAMPLES

Let  $E$  be a finite set, and let  $\mathcal{P}$  and  $\mathcal{K}$  be two families of subsets of  $E$ . We call  $(E, \mathcal{P}, \mathcal{K})$  a blocking system (on  $E$ ) if the following two axioms are satisfied:

- (1.1) For any partition of  $E$  into two sets  $E_1$  and  $E_0$  ( $E_0 \cap E_1 = \emptyset$  and  $E_0 \cup E_1 = E$ ), there is either a member of  $\mathcal{P}$  contained in  $E_1$  or a member of  $\mathcal{K}$  contained in  $E_0$ , but not both.
- (1.2) No member of  $\mathcal{P}$  contains another member of  $\mathcal{P}$ ; no member of  $\mathcal{K}$  contains another member of  $\mathcal{K}$ .

The first axiom (1.1) can be phrased in terms of painting elements of  $E$  with two colors: For any blue-red painting of  $E$ , there is either a blue  $P$  in  $\mathcal{P}$  or a red  $K$  in  $\mathcal{K}$ , but not both. The second axiom (1.2) is more a convenience than a necessity for our purposes, as will be clearer later on.

Observe that if  $(E, \mathcal{P}, \mathcal{K})$  is a blocking system, then for each  $P \in \mathcal{P}$  and  $K \in \mathcal{K}$ , we have  $P \cap K \neq \emptyset$ , by virtue of the last phrase in (1.1). In other words, each member of  $\mathcal{K}$  blocks all members of  $\mathcal{P}$ , and vice-versa. Note also that the axioms (1.1) and (1.2) are self-dual: Interchanging the roles of  $\mathcal{P}$  and  $\mathcal{K}$  alters neither (1.1) nor (1.2).

If  $\mathcal{P}$  is empty, then  $\mathcal{K} = \{\emptyset\}$  satisfies (1.1) and (1.2).

Examples of blocking systems abound. Some reasonably interesting ones will be described. But first we state and prove a theorem that indicates the great profusion of blocking systems. Its proof provides another characterization of blocking systems.

Following [7], we shall call a family  $\mathcal{P}$  of subsets of  $E$  a clutter on  $E$  if no member of  $\mathcal{P}$  contains another member of  $\mathcal{P}$ .

Theorem 1.1. Let  $E$  be a finite set and let  $\mathcal{P}$  be a clutter on  $E$ . Then there exists a unique clutter  $\mathcal{K}$  on  $E$  such that  $(E, \mathcal{P}, \mathcal{K})$  is a blocking system.

Proof. Let  $K \in \mathcal{K}$  if and only if  $K \cap P \neq \emptyset$  for all  $P \in \mathcal{P}$  and  $K$  is minimal with respect to this property. To verify that  $(E, \mathcal{P}, \mathcal{K})$  is a blocking system, it suffices to check (1.1). Thus consider a blue-red painting of  $E$ . Suppose there is no blue  $P \in \mathcal{P}$ . Let  $R$  be the set of all red members of  $E$  that belong to some  $P \in \mathcal{P}$ . Since there is no blue  $P \in \mathcal{P}$ , we have  $R \cap P \neq \emptyset$  for every  $P \in \mathcal{P}$ . Hence there is a  $K \in \mathcal{K}$  such that  $K \subseteq R$ , i.e., there is a red  $K \in \mathcal{K}$ . If there were both a blue  $P \in \mathcal{P}$  and a red  $K \in \mathcal{K}$ , then  $P \cap K = \emptyset$ , contradicting the definition of  $\mathcal{K}$ . Thus (1.1) holds and  $(E, \mathcal{P}, \mathcal{K})$  is a blocking system.

To establish uniqueness, let  $(E, \mathcal{P}, \mathcal{K})$  and  $(E, \mathcal{P}, \mathcal{K}')$  be blocking systems on  $E$  with  $\mathcal{K} \neq \mathcal{K}'$ . Interchanging the roles of  $\mathcal{K}$  and  $\mathcal{K}'$  if necessary, we may suppose  $K \in \mathcal{K} - \mathcal{K}'$ .

Consider the partition  $E - K, K$  of  $E$ . By (1.1) applied to  $(E, \mathcal{P}, \mathcal{K})$ , no subset of  $E - K$  is a member of  $\mathcal{P}$ . Hence by (1.1) applied to  $(E, \mathcal{P}, \mathcal{K}')$ , there is a  $K' \in \mathcal{K}'$  with  $K' \subset K$ . Now consider the partition  $E - K', K'$  of  $E$ . By (1.2), no subset of  $K'$  is a member of  $\mathcal{K}$ . Hence by (1.1) applied to  $(E, \mathcal{P}, \mathcal{K})$ , there is a  $P' \in \mathcal{P}$  with  $P' \subset E - K'$ . But then  $P'$  and  $K'$  violate (1.1) for the blocking system  $(E, \mathcal{P}, \mathcal{K}')$  and the partition  $E - K', K'$  of  $E$ . This contradiction proves Theorem 1.1.

Thus if  $\mathcal{P}$  is an arbitrary clutter on  $E$ , the family  $\mathcal{K} = \mathcal{P}^*$  of all "minimal blockers" of  $\mathcal{P}$  is the unique family of Theorem 1.1, and  $\mathcal{K}^* = \mathcal{P}^{**} = \mathcal{P}$ .

The primary role of (1.2) is to obtain uniqueness in Theorem 1.1. Uniqueness could be achieved in other ways. For instance, instead of normalizing to clutters  $\mathcal{P}$  and  $\mathcal{K}$  in Theorem 1.1, we could normalize to the families  $\mathcal{P}^+$  and  $\mathcal{K}^+$  of all supersets of members of  $\mathcal{P}$  and  $\mathcal{K}$ , respectively.

Some examples of blocking systems follow.

Example 1. Let  $E$  be the set of edges of a graph  $G$ ,  $\mathcal{P}$  the family of elementary paths joining two vertices of  $G$ , and  $\mathcal{K}$  the family of elementary cuts separating the two vertices.

Example 2. Let  $E$  be the set of cells in an  $n$  by  $n$  array; let  $\mathcal{P}$  be the family of subsets  $P \subset E$  having the property that there is just one cell of  $P$  in each row and column of the array; let  $\mathcal{K}$  be the family of subsets  $K \subset E$

such that  $K$  is a  $p$  by  $q$  subarray with  $p + q = n + 1$ .

(That  $(E, \mathcal{P}, \mathcal{K})$  is a blocking system follows from a well-known theorem of König [20] which asserts that in an  $n$  by  $n$   $(0, 1)$ -matrix, the maximum number of 1's, no two of which lie in the same row or column, is equal to the minimum number of rows and columns that contain all the 1's of the array.) More generally, let  $\mathcal{P}$  be the family of subsets  $P \subset E$  such that  $|P| = t$  and  $P$  has at most one cell in each row and column. Then  $\mathcal{K}$  is the family of subsets  $K \subset E$  such that  $K$  is a  $p$  by  $q$  subarray with  $p + q = 2n - t + 1$ .

Example 3. Let  $E = \{1, 2, \dots, 2k-1\}$ , let  $\mathcal{P}$  be the family of all  $k$ -element subsets of  $E$ , and let  $\mathcal{K} = \mathcal{P}$ . (In multi-person game theory, this example is known as the "straight majority game.")

Example 4. Let  $E$  be the set of edges of a graph  $G$ , let  $\mathcal{P}$  be the family of maximal trees of  $G$ , and let  $\mathcal{K}$  be the set of all elementary cuts (cocircuits) of  $G$ . (A tree of  $G$  is a subgraph of  $G$  that contains no circuit; a maximal tree is a tree of  $G$  that is maximal with respect to this property.)

Example 5. Let  $E$  be the set of edges of a graph  $G$ , let  $\mathcal{P}$  be the family of circuits in  $G$ , and let  $\mathcal{K}$  be the set of cotrees (complements in  $E$  of trees) of  $G$ .

Example 6. Let  $E'$  be the set of edges of a matroid  $(E', \mathcal{C})$ , let  $E = E' - \{e\}$  for some  $e \in E'$ , and let  $\mathcal{P}$  be the family of subsets  $P$  of  $E$  such that  $\{e\} \cup P \in \mathcal{C}$ .

Then  $\mathcal{K}$  is the family of subsets  $K$  of  $E$  such that  $\{e\} \cup K \in \mathcal{C}^*$ . Here  $(E', \mathcal{C}^*)$  is the matroid dual to  $(E', \mathcal{C})$ .

Example 7. Let  $E$  be the set of vertices of a graph  $G$ , and let  $\mathcal{P}$  be the family of pairs of adjacent vertices of  $G$  (two vertices are adjacent if they are joined by an edge.) Then  $\mathcal{K}$  is the family of subsets of vertices  $K$  such that  $K$  covers all edges of  $G$ , and is minimal with respect to this property. (In other words,  $\mathcal{K}$  is the family of all "minimal blockers" of  $\mathcal{P}$ .)

It is frequently difficult, as illustrated by Example 7, to find a useful description of the dual clutter  $\mathcal{K}$  of a simply described clutter  $\mathcal{P}$ .

One of the most important problems concerning blocking systems, a problem that arises time and again in applications, is the minimum covering or blocking problem: Given a simple description of  $\mathcal{P}$ , find a good algorithm that constructs  $K \in \mathcal{K}$  such that  $|K|$  is a minimum. For example, we might be given  $\mathcal{P}$  explicitly, say in the form of an incidence matrix  $A = (a(P, e))$ , where  $a(P, e) = 1$  or  $0$  according as  $e \in P$  or  $e \notin P$ . The minimum blocking problem then is equivalent to solving the following linear program in integers  $x(e) = 0$  or  $1$ :

$$(1.3) \quad \sum_{e \in E} a(P, e)x(e) \geq 1, \quad \text{all } P \in \mathcal{P},$$

$$\text{minimize } \sum_{e \in E} x(e).$$

Various methods have been proposed for such problems, but no good algorithms are known. Indeed, most of the methods that have been proposed can be shown to be bad: the amount of computational effort increases exponentially with the size of the problem.

There is a good algorithm, however, for computing the following lower bound on the minimum in (1.3). Consider the class  $\mathcal{Q}$  of all  $(0, 1)$ -matrices having the same row and column sums as  $A$ . For  $A$  in  $\mathcal{Q}$ , let  $w(A)$  denote the minimum in (1.3), and let

$$(1.4) \quad \tilde{w} = \min_{A \in \mathcal{Q}} w(A).$$

The integer  $\tilde{w}$  has been explicitly evaluated by Fulkerson and Ryser in [14], and a very simple construction for a matrix  $\tilde{A}$  in  $\mathcal{Q}$  such that  $w(\tilde{A}) = \tilde{w}$  has been given in [15].

## 2. THE MIN-MAX THEOREM

The analogue of Theorem 3.1, Part I, is valid for all blocking systems, and can be viewed as characterizing blocking systems:

Theorem 2.1. Let  $(E, \mathcal{P}, \mathcal{K})$  be a blocking system, and let  $f$  be a real-valued function defined on  $E$ . Then

$$(2.1) \quad \max_{P \in \mathcal{P}} \min_{e \in P} f(e) = \min_{K \in \mathcal{K}} \max_{e \in K} f(e).$$

Conversely, if  $\mathcal{P}$  and  $\mathcal{K}$  are clutters on  $E$  such that (2.1)

holds for every real-valued  $f$  defined on  $E$ , then  $(E, \mathcal{P}, \mathcal{K})$  is a blocking system.

Proof. The proof that (2.1) holds for a blocking system is entirely analogous to the proof of Theorem 3.1, Part I. In brief: The left-hand side of (2.1) is less than or equal to the right-hand side since  $P \cap K$  is nonempty for each  $P \in \mathcal{P}$ ,  $K \in \mathcal{K}$ . To establish equality, order the elements of  $E$  according to decreasing values of  $f$ ; then paint elements of  $E$  blue, one after another, until the blue set first contains an element of  $\mathcal{P}$ .

(In other words, the threshold method establishes equality in (2.1) and simultaneously evaluates (2.1). It will be a good method for this evaluation in case there is a good method for recognizing whether an arbitrary subset of  $E$  contains a member of  $\mathcal{P}$  (or a member of  $\mathcal{K}$ ).)

Conversely, let  $\mathcal{P}$  and  $\mathcal{K}$  be clutters on  $E$  and suppose (2.1) holds for every real-valued  $f$  defined on  $E$ . Let  $f(e) = 1$  or  $0$  according as  $e$  is blue or red. Suppose there is no blue  $P \in \mathcal{P}$ . Then

$$\max_{P \in \mathcal{P}} \min_{e \in P} f(e) = 0 = \min_{K \in \mathcal{K}} \max_{e \in K} f(e).$$

If there were no red  $K \in \mathcal{K}$ , we would have

$$\min_{K \in \mathcal{K}} \max_{e \in K} f(e) = 1,$$

a contradiction. Hence there is a red  $K \in \mathcal{K}$ . On the other hand, if there were both a blue  $P \in \mathcal{P}$  and a red  $K \in \mathcal{K}$ , then

$$\max_{P \in \mathcal{P}} \min_{e \in P} f(e) = 1, \min_{K \in \mathcal{K}} \max_{e \in K} f(e) = 0,$$

contradicting (2.1). Hence  $(E, \mathcal{P}, \mathcal{K})$  is a blocking system.

### 3. THE LENGTH-WIDTH INEQUALITY AND MAX-FLOW MIN-CUT EQUALITY

Let  $(E, \mathcal{P}, \mathcal{K})$  be a blocking system, and suppose  $\iota(e)$ ,  $w(e)$  are two nonnegative numbers associated with element  $e \in E$ . Define the length of the system to be

$$(3.1) \quad \lambda = \min_{P \in \mathcal{P}} \sum_{e \in P} \iota(e),$$

and the width to be

$$(3.2) \quad \omega = \min_{K \in \mathcal{K}} \sum_{e \in K} w(e).$$

Following Lehman [22], we shall say that the length-width inequality holds for  $(E, \mathcal{P}, \mathcal{K})$  if

$$(3.3) \quad \lambda \omega \leq \sum_{e \in E} \iota(e) w(e)$$

is satisfied for every pair of nonnegative functions  $\iota$ ,  $w$  defined on  $E$ .

For instance, if  $(E, \mathcal{P}, \kappa)$  is the blocking system of Example 1, we have seen in Part I that the length-width inequality holds. It also holds for Example 6 provided the underlying matroid is regular; this is a corollary of Theorem 3.4, Part II. On the other hand, the length-width inequality fails for the blocking system of Example 3.

For each  $P \in \mathcal{P}$  and  $e \in E$ , define  $a(P, e) = 1$  or  $0$  according as  $e \in P$  or  $e \notin P$ . Now consider the linear program

$$(3.3) \quad \sum_{P \in \mathcal{P}} y(P) a(P, e) \leq w(e), \quad e \in E,$$

$$y(P) \geq 0, \quad P \in \mathcal{P},$$

$$\text{maximize } \sum_{P \in \mathcal{P}} y(P).$$

Clearly the maximum in (3.3) is less than or equal to the width of  $(E, \mathcal{P}, \kappa)$ . If equality holds here for every nonnegative  $w$  defined on  $E$ , we say, as in [22], that the max-flow min-cut equality holds for  $(E, \mathcal{P}, \kappa)$ .

Thus, for instance, the max-flow min-cut equality holds for Example 1, for Example 6 if the underlying matroid is regular, and fails for Example 3, just as for the length-width inequality. This behavior is not accidental. One of the main results of [22] is that the max-flow min-cut equality holds for a blocking system

if and only if the length-width inequality holds.

Consequently, if the max-flow min-cut equality holds for  $(E, \mathcal{P}, \mathcal{K})$ , it also holds for  $(E, \mathcal{K}, \mathcal{P})$ , since the roles of  $\mathcal{P}$  and  $\mathcal{K}$  are symmetric in the length-width inequality.

In any event, the problem of evaluating the width of a blocking system for a given nonnegative function  $w$  is a generalization of the minimum blocking problem mentioned earlier. It would be interesting to discover other significant classes of blocking systems for which the length-width inequality, and hence the max-flow min-cut equality, holds.

BIBLIOGRAPHY

1. Berge, C. and A. Ghouila-Houri, Programmes, Jeux et Réseaux de Transport, Dunod, Paris, 1962.
2. Camion, P., "Application d'une généralisation du lemme de Minty à un problème d'infimum de fonction convexe," Cahiers Centre Études Recherche Opér., 7, 1965, 230-247.
3. Dantzig, G. B., Linear Programming and Extensions, Princeton University Press, Princeton, New Jersey, 1963.
4. Dantzig, G. B., and D. R. Fulkerson, "On the Max-Flow Min-Cut Theorem of Networks, Linear Inequalities and Related Systems, Ann. of Math. Study 38, Princeton University Press, Princeton, New Jersey, 1956, 215-221.
5. Duffin, R. J., "The Extremal Length of a Network," J. Math. Anal. and Appl., 5, 1962, 200-215.
6. Edmonds, J., "Paths, Trees, and Flowers," Canadian J. Math., 17, 1965, 449-467.
7. Edmonds, J. and D. R. Fulkerson, Bottleneck Extrema (to appear).
8. Ford, L. R., Jr. and D. R. Fulkerson, Flows in Networks, Princeton University Press, Princeton, New Jersey, 1962.
9. ———, "Maximal Flow Through a Network," Canadian J. Math., 1956, 399-404.
10. ———, "A Simple Algorithm for Maximal Network Flows and an Application to the Hitchcock Problem," Canadian J. Math., 9, 1957, 210-218.
11. ———, "Constructing Maximal Dynamic Flows from Static Flows," Op. Res., 6, 1958, 419-433.
12. Fulkerson, D. R., "An Out-of-Kilter Method for Minimal Cost Flow Problems," J. Soc. Indust. Appl. Math., 9, 1961, 18-27.
13. ———, "Flow Networks and Combinatorial Operations Research," Amer. Math. Monthly, 73, 1966, 115-138.
14. Fulkerson, D. R. and H. J. Ryser, "Widths and Heights of  $(0, 1)$ -Matrices," Canadian J. Math., 13, 1961, 239-255.
15. ———, "Multiplicities and Minimal Widths for  $(0, 1)$ -Matrices," Canadian J. Math., 14, 498-508.

16. Gale, D., H. W. Kuhn, and A. W. Tucker, "Linear Programming and the Theory of Games," Activity Analysis of Production and Allocation, John Wiley and Sons, New York, 1951, 317-329.
17. Gross, O., The Bottleneck Assignment Problem, The RAND Corporation, P-1630, 1959.
18. Hoffman, A. J., "Some Recent Applications of the Theory of Linear Inequalities to Extremal Combinatorial Analysis," Proc. Symposia Applied Math., 10, 1960.
19. Hu, T. C., "The Maximum Capacity Route Problem," Op. Res., 9, 1961, 898-900.
20. Konig, D., Theorie der Endlichen und Unendlichen Graphen, Chelsea, New York, 1950.
21. Lawler, E., "Covering Problems: Duality Relations and a New Method of Solution," SIAM Journal, 14, 1966, 1115-1133.
22. Lehman, A., On the Width Length Inequality (mimeographed).
23. Minty, G. J., "Monotone Networks," Proc. Roy. Soc., A, 257, 1960, 194-212.
24. ———, "On the Axiomatic Foundations of the Theories of Directed Linear Graphs, Electrical Networks, and Network-Programming," J. Math. Mech. 15, 1966, 485-520.
25. Moore, E. F. and C. E. Shannon, "Reliable Circuits Using Less Reliable Relays," J. Franklin Inst. 262, 1956, 191-208.
26. Pollack, M., "The Maximum Capacity Route Through a Network," Op. Res. 8, 1960, 733-736.
27. Robacker, J. T., Min-Max Theorems on Shortest Chains and Disjunct Cuts of a Network, The RAND Corporation, RM-1660-PR, 1956.
28. Rockafellar, R. T., "The Elementary Vectors of a Subspace of  $R^N$ ," to appear in the Proceedings of the Chapel Hill Symposium on Combinatorial Mathematics and Its Applications.
29. Shapley, L. S., "Simple Games: An Outline of the Descriptive Theory," Behavioral Sciences, 7, 1962, 59-66.
30. Tutte, W. T., "A Class of Abelian Groups," Canadian J. Math., 8, 1956, 13-28.

31. \_\_\_\_\_, "A Homotopy Theorem for Matroids I, II,"  
Trans. Amer. Math. Soc., 88, 1958, 144-174.
32. \_\_\_\_\_, "Matroids and Graphs," Trans. Amer. Math.  
Soc., 90, 1959, 527-552.
33. \_\_\_\_\_, "Lectures on Matroids," J. Res. Nat. Bur.  
Std., B, 69, 1965, 1-47.
34. \_\_\_\_\_, Introduction to the Theory of Matroids,  
The RAND Corporation, R-448-PR, 1966.
35. Whitney, H., "On the Abstract Properties of Linear  
Dependence," Amer. J. Math., 57, 1935, 507-553.

Lectures on  
**COMBINATORIAL METHODS**

by  
**JACK EDMONDS**

at the  
**American Mathematical Society Summer Seminar**  
on the  
**Mathematics of the Decision Sciences**  
**Stanford University**  
**July - August 1967**

OPTIMUM BRANCHINGS

Jack Edmonds  
National Bureau of Standards  
Washington, D.C.

# Optimum Branchings<sup>\*</sup>

by

Jack Edmonds

National Bureau of Standards

Washington, D.C.

To Professor Marcel Riesz on his 80th birthday

§1

A (directed) graph  $G$ , for purposes here, is a finite set of nodes and a finite set of edges, where each edge is said to be directed toward one of the nodes, called the front end of the edge, and said to be directed away from a different one of the nodes, called the rear end of the edge. An edge and each of its ends are said to meet. A subgraph of  $G$  is a subcollection of its members which, under the same incidence relations, is a graph. A graph is called connected if it is not empty and its members do not partition into two disjoint non-empty subgraphs. A polygon is a connected graph  $Q$  such that each node of  $Q$  meets exactly two edges of  $Q$ . An (elementary uniformly directed) circuit is a polygon which contains one edge directed toward, and one edge directed away from, each of its nodes. A forest is a graph which contains no polygon. A tree is a connected forest. A branching is a forest whose edges are directed so that each is directed toward a different

---

<sup>\*</sup>Prepared while the author was a visiting professor at the University of Waterloo, Ontario, Canada. Presented under the title Optimum Arborescences at the International Seminar on Graph Theory and its Applications, Rome, July 1966.

node. An arborescence is a connected branching. An (elementary uniformly directed) path  $P$  is an arborescence such that each edge in  $P$  is directed away from a different node, and such that there is at least one edge in  $P$ .

We shall occasionally use "obvious" facts about graphs without justifying them.

Clearly, a branching (forest) is the union of a unique family of disjoint arborescences (trees).

Exactly one node in an arborescence  $T$ , called the root of  $T$ , has no edge of  $T$  directed toward it. A branching (forest) is an arborescence (tree) if and only if it has exactly one less edge than nodes. No branching (forest) has more edges than this.

In a path  $P$  there are exactly two nodes, called the ends of  $P$ , which each meet only one edge in  $P$ . The rest of the nodes in  $P$  each meet exactly two edges in  $P$ . A path  $P$  is said to go from the node which is only a rear end in  $P$  (the root of  $P$ ) to the node which is only a front end in  $P$ . For any arborescence  $T$ , and any node  $v$  in  $T$  except the root, there is a unique path in  $T$  going from the root to  $v$ . Any path in  $T$  going to  $v$  and any path in  $T$  going from  $v$  have only  $v$  in common, and their union is a path. And so on.

§2

Let  $G$  be any graph with a real numerical weight  $c_j$  corresponding to each edge  $e_j \in G$ . The problem treated here is to find in  $G$  a branching  $B$  which has maximum total weight,  $\sum c_j$ , summed over  $e_j \in B$ .  $B$  is called an optimum branching in  $G$ .

First we show that certain variations of the problem reduce immediately to it.

A spanning subgraph of  $G$  is a subgraph which contains all the nodes of  $G$ . A branching in  $G$  is a spanning arborescence of  $G$  if and only if the number of its edges is one less than the number of nodes in  $G$ . No branching in  $G$  can have more edges than this.

An optimum branching in  $G$  of course contains no edge with negative weight, and indeed may be empty if all  $c_j \leq 0$ . Even if all  $c_j > 0$  and  $G$  contains a spanning arborescence, an optimum branching in  $G$  need not be an arborescence.

If there is a spanning arborescence  $T$  in  $G$ , then an optimum one, i.e. one which has maximum total weight,  $\sum c_j$ ,  $e_j \in T$ , can be found as an optimum branching in  $G$  where the edges carry new weights  $c_j' = c_j + h$ ,  $h > \sum |c_j|$ ,  $e_j \in G$ . A spanning arborescence in  $G$  which is optimum relative to weights  $c_j$ ,  $e_j \in G$ , is also optimum relative to weights  $c_j + k$ ,  $e_j \in G$ , for any constant  $k$ , since every spanning arborescence has the same number of edges.

Constant  $h$  is larger than the difference in total weights (relative to weights  $c_j, e_j \in G$ ) of any two branchings in  $G$ . It follows that an optimum branching in  $G$ , relative to weights  $c_j' = c_j + h$ , will be a branching with a maximum number of edges. In particular, it will be a spanning arborescence if and only if  $G$  contains a spanning arborescence.

A spanning arborescence  $T$  in  $G$  which has minimum total weight,  $\sum c_j, e_j \in T$ , is the same as one which has maximum total weight  $\sum c_j', e_j \in T$ , relative to weights  $c_j' = -c_j$ .

It will be evident that the efficiency of the method for treating optimum branchings is not seriously effected by a large change  $h$  (say of the form  $10^n$ ) in all the weights. In fact the method is easily modified to treat optimum spanning arborescences directly.

If there is a spanning arborescence in  $G$  which is rooted at a prescribed node, say  $r$ , then an optimum one can be found by finding an optimum spanning arborescence in the graph  $G'$  obtained from  $G$  by adjoining a new edge  $e_0$  (carrying arbitrary weight  $c_0$ ) which is directed toward  $r$  and directed from a new node having no other incident edges. Clearly,  $T$  is a spanning arborescence in  $G$  which is rooted at  $r$  if and only if  $T$  together with  $e_0$  is a spanning arborescence of  $G'$ .

If the edges in graph  $G$  represent the links for possible direct communication from one node to another, then each  $c_j$  is the cost of direct communication from the rear end of  $e_j$  to the front end of  $e_j$ , and if cost is additive,

then a minimum-total-weight spanning arborescence rooted at prescribed node  $r$  represents the least costly way to have a message communicated from  $r$  to all other nodes of  $G$ .

Another application is where it is desired to arrange an institution into an optimum heirarchy (branchocracy).

### §3

Our main result is

Theorem 1. There exists a good algorithm for finding, in any graph  $G$  with a numerical weight corresponding to each edge, an optimum branching.

We say an algorithm is good if there is a polynomial function  $f(n)$  which, for every positive-integer valued  $n$ , is an upper bound on the "amount of work" the algorithm does for any input of "size"  $n$ . The concept is easy to formalize — — relative, say, to a Turing machine, or relative to any typical digital computer with an unlimited supply of tape.

For optimum branching, the largest number of significant digits in an edge weight, as well as the number of edges of  $G$ , must be figured somehow into the measure  $n$  of input "size". One might for example take  $n$  to be the maximum of these two numbers or to be the vector consisting of both numbers.

theorem

The proof of Theorem 1 is constructive. The/is proved by displaying one particular algorithm for optimum branching which is obviously good.

If we remove from the optimum-spanning-arborescence problem the condition that each member of the set  $T$  of edges being optimized must have a different front end, then we get the optimum-spanning-tree problem. That is to find, if there is one, in any graph  $G$  with a numerical weight on each edge, a spanning tree which has maximum (or minimum) total weight.

Especially simple algorithms are well-known for this problem [cf. 5 and 6]. One is, starting with an empty bucket, build up a set of elements having "admissible structure" by putting elements into the bucket one after another as long as possible, so that each addition is a maximum weight element among those not in the bucket which, together with the ones already in the bucket, would preserve admissible structure. For the optimum-spanning-tree problem, the elements are the edges of  $G$  and "admissible" means "forest". The algorithm is certainly good. It is also valid for that problem.

Where admissible" means "branching", the above algorithm is not generally valid for finding an optimum spanning arborescence. Paper [3] abstractly characterizes those structures for which this "greedy algorithm" is valid for any numerical weighting.

If we add to the conditions of the optimum-spanning-arborescence problem the condition that each member of the set of edges being optimized is to have a different rear end, then we have the problem of finding, if there is one, an optimum spanning (uniformly directed) path in any graph  $G$  with a numerical weight on each edge. This is a version of the well-known traveling salesman problem [cf. 4]. I conjecture that there is no good algorithm for the traveling

saleman problem. My reasons are the same as for any mathematical conjecture:

(1) It is a legitimate mathematical possibility, and (2) I do not know.

A matching in a graph is a subset of its edges such that no two of them meet the same node. A good algorithm is known for finding, in any graph with a numerical weight on each edge, a maximum-total-weight matching. The treatment [1 and 2] of maximum matchings and the treatment here of optimum branchings are similar, though the structural details are different and maximum matching is more complicated.

§4

Here is the algorithm for finding a maximum-total weight branching in any (directed) graph  $G$  with a numerical weight  $c_j$  on each edge  $e_j \in G$ . Recall that a branching is a forest such that each edge is directed toward a different node.

Begin the algorithm by applying instruction (I1) where  $G^1$  is  $G^0 = G$  and where  $D^1$  and  $E^1$  are empty buckets,  $D^0$  and  $E^0$ .

(I1) Choose a node  $v$  in  $G^1$  and not in  $D^1$ . Put  $v$  into bucket  $D^1$ . If there is in  $G^1$  a positively weighted edge directed toward  $v$ , put one of them having maximum weight into bucket  $E^1$ .

Repeat (I1) until

- (a)  $E^i$  no longer comprises the edges of a branching in  $G^i$ , or until
- (b) every node of  $G^i$  is in  $D^i$ , and  $E^i$  does comprise the edges of a branching. When case (a) occurs, apply (I2).

For convenience assume that every branching which we consider in graph  $G^i$  contains all the nodes of  $G^i$ . We say that a set of edges in  $G^i$  forms the unique subgraph of  $G^i$  consisting of those edges and all nodes in  $G^i$ .

Each edge  $e$  put into  $E^i$  according to (I1) is directed toward a node  $v$  which is the root of a connected component of the branching, say  $B$ , formed by the edges in  $E^i$  before  $e$  is put into  $E^i$ . If the rear end  $v_6$  of  $e$  is in a different component of  $B$  than  $v$ , then  $B \cup e$  is a branching, and so when  $e$  is put into  $E^i$ , (a) does not hold.

If  $v_6$  is in the same component of  $B$  as  $v$ , then  $B$  contains a unique path  $P$  going from  $v$  to  $v_6$ . In this case,  $Q^i = P \cup e$  is a circuit contained in  $B \cup e$ , so as soon as  $e$  is put into  $E^i$ , (a) does hold.

(I2) Store  $Q^i$  and a specification of one of the edges, say  $e_0^i$ , of  $Q^i$  which has minimum weight in  $Q^i$  relative to the edge-weights for  $G^i$ . Obtain a new graph  $G^{i+1}$  from  $G^i$  by "shrinking" to a single new node,  $v_1^{i+1}$ , the circuit  $Q^i$  and every edge of  $G^i$  which has both ends in  $Q^i$ . The edges (denoted as  $e_j^{i+1}$ ) of  $G^{i+1}$  are those edges (denoted as  $e_j^i$ ) of  $G^i$  which have at most one end in  $Q^i$ . Every edge of  $G^i$  which has one end in  $Q^i$  will in  $G^{i+1}$  have  $v_1^{i+1}$  at that end. All other edge-ends are the same in  $G^{i+1}$  as in  $G^i$ . The nodes of  $Q^i$  are not in  $G^{i+1}$ .

Every edge, say  $e_3^{i+1}$ , which as  $e_3^i$  in  $G^i$  is directed toward a node, say  $v_3^i$ , in  $Q^i$  and directed away from a node not in  $Q^i$ , gets a possibly different weight for  $G^{i+1}$ :

$$(1) \quad c_3^{i+1} = c_3^i + c_0^i - c_4^i$$

where  $c_3^i$  is the weight of  $e_3^i$  for  $G^i$ , where  $c_0^i$  is the minimum weight for  $G^i$  of an edge, say  $e_0^i$ , in  $Q^i$ ; and where  $c_4^i$  is the weight for  $G^i$  of the unique edge, say  $e_4^i$ , which is in  $Q^i$  and directed toward  $v_3^i$ .

All other edges in  $G^{i+1}$  keep the same weight as for  $G^i$ .

In justifying the algorithm we shall make use of the following relations

$$(2) \quad c_0^i \geq 0, \quad (3) \quad c_4^i \geq c_0^i, \quad \text{and} \quad (4) \quad c_4^i \geq c_3^i.$$

Put into bucket  $D^{i+1}$  the nodes which are in both  $G^{i+1}$  and bucket  $D^i$ . Put into bucket  $E^{i+1}$  the edges which are in both  $G^{i+1}$  and bucket  $E^i$ , i.e., put into bucket  $E^{i+1}$  the final contents of bucket  $E^i$  minus the edges of circuit  $Q^i$ . It is easy to see that the edges in bucket  $E^{i+1}$  form a branching in  $G^{i+1}$ .

Continue the algorithm by

applying (I 1) where  $i$  is one greater.

Eventually, after a small number of applications of (I 1) and (I 2), case (b) must occur.

As soon as (b) occurs, for say  $i = k$ , (I 1) and (I 2) are never applied again. Instead, (I 3) is applied successively for  $i + 1 = k, k-1, \dots, 1$ , until the graph  $G^1$  obtained is the original  $G$ . At that point, the branching  $B^1 = B^0$  is a maximum-total-weight branching of  $G$ .

The final contents of bucket  $E^k$  form a branching in graph  $G^k$  which we call  $B^k$ .

(I 3) It is not difficult to see that since  $B^{i+1}$  is a forest in  $G^{i+1}$  and since  $G^{i+1}$  is obtained from  $G^i$  by shrinking the circuit  $Q^i$  in  $G^i$  (and all edges of  $G^i$  with both ends in  $Q^i$ ) to the node  $\mu_1^{i+1}$  of  $G^{i+1}$ , the subgraph  $H^i$  of  $G^i$ , formed by the edges in  $B^{i+1}$  and the edges in  $Q^i$ , contains only one polygon, namely  $Q^i$ .

In the case where  $\mu_1^{i+1}$  is not a root of (a connected component of) branching  $B^{i+1}$  in  $G^{i+1}$ , there is a unique edge, say  $e_1^{i+1}$ , of  $B^{i+1}$  which is directed toward  $\mu_1^{i+1}$ . In  $G^i$ ,  $e_1^i$  is directed toward a node, say  $\mu_2^i$ , of  $Q^i$ . Since  $Q^i$  is a circuit, there is a unique edge, say  $e_2^i$ , of  $Q^i$  which is directed toward  $\mu_2^i$ . Clearly,  $e_1^i$  and  $e_2^i$  are the only two edges of  $H^i$  which are directed toward the same node. Thus, since  $e_2^i$  is in the only polygon of  $H^i$ , deleting  $e_2^i$  from  $H^i$  yields a branching in  $G^i$ , which is called  $B^i$ .

In the case where  $\mu_1^{i+1}$  is a root of branching  $B^{i+1}$  in  $G^{i+1}$ , i.e., where no edge of  $B^{i+1}$  is directed toward  $\mu_1^{i+1}$ , no two edges of  $H^i$  are directed toward the same node. Therefore, deleting any edge of  $Q^i$  from  $H^i$  yields a branching in  $G^i$ . To obtain the branching  $B^i$  in  $G^i$ , delete from  $H^i$  one of the edges  $e_0^i$  of  $Q^i$  which has minimum weight  $c_0^i$ .

That completes the description of the algorithm. Evidently it is a good algorithm. Evidently its output is a branching  $B^0$  in graph  $G$ . In order to prove Theorem 1, what remains to be done is prove that  $B^0$  has maximum total weight.

§5

Theorem 1 and the following geometric theorem are proven together.

Let  $G$  be any graph. (No edge-weights are specified.) Let there be a real variable  $x_j$  for each edge  $e_j \in G$ . Let  $P_G$  be the polyhedron of vectors  $x = [x_j]$  which satisfy the system  $L_G$ , consisting of inequalities  $L_1$ ,  $L_2$ , and  $L_3$ .

( $L_1$ ) For every edge  $e_j \in G$ ,  $x_j \geq 0$ .

( $L_2$ ) For every node  $v \in G$ ,  $\sum x_j \leq 1$ , summed over all  $j$ 's such that  $e_j$  is directed toward  $v$ .

( $L_3$ ) For every set  $S$  of two or more nodes in  $G$ ,  $\sum x_j \leq |S| - 1$ , summed over all  $j$ 's such that  $e_j$  has both ends in  $S$ . ( $|S|$  denotes the cardinality of  $S$ .)

Any vector  $x = [x_j]$  of zeroes and ones is called the (incidence) vector of the subset of  $e_j$ 's such that  $x_j = 1$ .

Theorem 2. The vertices of polyhedron  $P_G$  are precisely the vectors of the subsets of edges in  $G$  which comprise branchings.

A polyhedron (convex polyhedron)  $P$  is the set of all the vectors, i.e., points, which satisfy some finite system  $L$  of linear inequalities. A vertex (extreme point) of  $P$  is a point which, for some linear function, is the unique point in  $P$  which maximizes that function.

A basic point  $x = x^0$  of a finite system  $L$  of linear inequalities is the unique solution of a system,  $\sum_i a_{ij} x_i = b_j, j \in J$ , such that  $\sum_i a_{ij} x_i \leq b_j, j \in J$ , is a subsystem of  $L$ .

If basic point  $x^0$  of  $L$  is in the polyhedron  $P$  of  $L$ , then it is a vertex of  $P$ , because clearly  $x^0$  is then the unique point in  $P$  which maximizes  $\sum_i (\sum_j a_{ij}) x_i, j \in J$ .

We shall see without difficulty that any point  $x^0$ , which is the vector of a branching say  $B^0$  in  $G$ , is a vertex of  $P_G$ . Vector  $x^0$  satisfies  $L_1$  since it is all zeroes and ones. Vector  $x^0$  satisfies  $L_2$  for any node  $y \in G$ , since, by the definition of branching, at most one of the  $x_j$ 's in this inequality has value 1 for  $x^0$ .

The branching  $B^0$  is a forest, so any set  $S$  of nodes, together with the subset  $E_S^0$  of the edges in  $B^0$  which have both ends in  $S$ , forms a forest. The number of edges in a forest is at most the number of nodes in the forest minus 1; in particular,  $|E_S^0| \leq |S| - 1$ . Therefore, vector  $x^0$  satisfies  $L_3$  for any subset  $S$  of (two or more) nodes in  $G$ , since  $|E_S^0|$  of the  $x_j$ 's in this inequality have the value 1 for  $x^0$ . Summarizing the conclusion so far,  $x^0$  is a point in  $P_G$ .

Vector  $x^0$  is the unique solution of the linear system:  $x_j = 0$  for every edge  $e_j$  not in  $B^0$ , and  $\sum x_j = 1$  (summed over  $e_j$ 's directed toward  $v$ ) for every node  $v$  which has some edge of  $B^0$  directed toward it. This system can be obtained from certain of the relations of  $L_1$  and  $L_2$  by replacing their inequality signs. Therefore  $x^0$  is a basic point of  $L_G$ , and hence a vertex of  $P_G$ .

Most of this paper is directed toward proving:

Lemma 1: Every linear function,  $\sum c_j x_j$  (summed over all edges  $e_j \in G$ ), is maximized in  $P_G$  by the vector of some branching in  $G$ .

From Lemma 1 and from the definition of vertex, it follows immediately that every vertex of  $P_G$  is the vector of a branching in  $G$ . This will conclude the proof of Theorem 2.

A branching  $B^0$  in graph  $G$  has maximum total weight relative to the vector  $c = [c_j]$  of edge-weights if and only if the vector  $x^0 = [x_j^0]$  of  $B^0$  maximizes  $(c, x) = \sum c_j x_j$  over all vectors of branchings in  $G$ . If  $x^0$  maximizes  $(c, x)$  over  $P_G$ , then it maximizes  $(c, x)$  over the vectors of branchings in  $G$ , since the latter are in  $P_G$ .

Our task, therefore, is to show that the vector of the branching  $B^0$ , produced by the algorithm, maximizes  $(c, x)$  over  $P_G$ . This will prove that the algorithm is valid and will prove Lemma 1.

§6

The following computations are well-known in linear programming.

Suppose that  $x = [x_\xi]$  is any vector which satisfies

$$(5) \quad x_\xi \geq 0 \quad \text{for every } \xi, \text{ and}$$

$$(6) \quad \sum_{\xi} a_{\xi\eta} x_\xi \leq b_\eta \quad \text{for every } \eta,$$

and that  $y = [y_\eta]$  is any vector which satisfies

$$(7) \quad y_\eta \geq 0 \quad \text{for every } \eta, \text{ and}$$

$$(8) \quad \sum_{\eta} a_{\xi\eta} y_\eta \geq c_\xi \quad \text{for every } \xi.$$

Since (6) and (7) imply

$$(9) \quad \sum_{\eta} \left( \sum_{\xi} a_{\xi\eta} x_\xi \right) y_\eta \leq \sum_{\eta} b_\eta y_\eta = (b, y),$$

and since (5) and (8) imply

$$(10) \quad \sum_{\xi} \left( \sum_{\eta} a_{\xi\eta} y_\eta \right) x_\xi \geq \sum_{\xi} c_\xi x_\xi = (c, x),$$

we have

$$(11) \quad (c, x) \leq (b, y).$$

Since (11) holds for any  $x$  and any  $y$ , if  $(c, x^0) = (b, y^0)$  holds for particular  $x = x^0$  and  $y = y^0$ , then  $x^0$  must maximize  $(c, x)$  and  $y^0$  must minimize  $(b, y)$ .

Suppose for particular  $x = x^1$  and  $y = y^1$  that

$$(12) \quad \xi \sum a_{\xi \eta} x_{\xi}^1 = b_{\eta} \text{ for } \eta \text{ such that } y_{\eta}^1 \neq 0,$$

and

$$(13) \quad \eta \sum a_{\xi \eta} y_{\eta}^1 = c_{\xi} \text{ for } \xi \text{ such that } x_{\xi}^1 \neq 0.$$

Since (12) implies equality in (9), and (13) implies equality in (10), we have  $(c, x^1) = (b, y^1)$ . Therefore,

$$(14) \quad \begin{aligned} x^1 & \text{ maximizes } (c, x) \text{ and} \\ y^1 & \text{ minimizes } (b, y). \end{aligned}$$

Our present interest is where (5) is  $(L_1)$ , and (6) is  $(L_2)$  and  $(L_3)$ . For any linear function  $(c, x) = \sum_j c_j x_j$  of points  $x \in P_G$ , we get a dual system, (7), (8),  $(b, y)$ , by letting a variable  $y_{\eta}$  correspond to each inequality of  $L_2$  and  $L_3$ . That is let a variable  $y_h$  correspond to each node  $u_h \in G$  and let a variable  $y_s$  correspond to each set  $S$  of two or more nodes in  $G$ .

For (7) we have,

$$(15) \quad \text{for every } v_h, y_h \geq 0, \text{ and}$$

$$(16) \quad \text{for every } S, y_s \geq 0.$$

Coefficient  $a_{jh} = 1$  if edge  $e_j$  is directed toward node  $v_h$ , and  $a_{jh} = 0$  otherwise. Coefficient  $a_{js} = 1$  if edge  $e_j$  has both ends in  $S$ , and  $a_{js} = 0$  otherwise. For every  $v_h$ ,  $b_h = 1$ . For every  $S$ ,  $b_s = |S| - 1$ .

Therefore, (8) becomes

$$(17) \quad \text{for every edge } e_j \in G, \\ y_h + w_j \geq c_j, \text{ where } v_h \text{ is the front end} \\ \text{of } e_j, \text{ and where } w_j = \sum y_s, \text{ summed over} \\ \text{all sets } S \text{ which contain both ends of } e_j.$$

Function  $(b, y)$  becomes

$$(b, y) = \sum_h y_h + \sum_s (|S| - 1) y_s,$$

summed over all  $v_h$  and over all  $S$ .

Recall that our task is to show that the vector  $x^0$  of the branching  $B^0$ , produced by the algorithm, maximizes  $(c, x)$  over  $P_G$ .

In view of (14), we do so by constructing a vector  $y = [y_h, y_s]$  which satisfies (15), (16), (17), and which satisfies (12) and (13). For the present system, (12) is

$$(18) \quad \text{for every node } v_h \text{ such that } y_h \neq 0, \\ \sum_j x_j^0 = 1, \text{ summed over } j\text{'s such that } e_j \\ \text{is directed toward } v_h; \text{ and}$$

- (19) for every set  $S$  such that  $y_s \neq 0$ ,  $\sum x_j^0 = |S|-1$ ,  
summed over  $j$ 's such that  $e_j$  has both ends in  $S$ .

In other words, (18) says that if  $y_h \neq 0$  then an edge of the branching  $B^0$  is directed toward  $v_h$ , and (19) says that if  $y_s \neq 0$  then exactly  $|S|-1$  edges of  $B^0$  have both ends in  $S$ .

For the present system, (13) is

- (20) for every edge  $e_j$  in the branching  $B^0$ ,  
 $y_h + w_j = c_j$ , where  $v_h$  and  $w_j$  are as in (17).

§7

For each graph  $G^i$  ( $i = k, k-1, \dots, 0$ ) with weight  $c_j^i$  on each edge  $e_j^i \in G^i$ , and for the branching  $B^i$  in  $G^i$ , we will describe a vector  $y^i$  which satisfies (15) - (20), where  $G$  and  $B^0$  are replaced by  $G^i$  and  $B^i$  and where vector  $y$  is  $y^i$ .

First we describe a  $y^k$ , and then, assuming a  $y^{i+1}$  ( $i = k-1, \dots, 0$ ), we describe a  $y^i$ . Thus by induction we obtain a  $y = y^0$  and the proof of Theorems 1 and 2.

The vector  $y^k = [y_h^k, y_s^k]$  is  $y_s^k = 0$  for every set  $S$  of two or more nodes in  $G^k$ ,  $y_h^k = 0$  for every node  $v_h^k$  in  $G^k$  which has no edge of  $B^k$  directed toward it, and, for every other node  $v_h^k$  in  $G^k$ ,  $y_h^k = c_j^k$  where

edge  $e_j^k$  of  $B^k$  is directed toward  $v_h^k$ . Conditions (15) - (20) for  $y^k$  can be immediately verified from the fact that for every node  $v_h^k \in G^k$  either there is no edge of  $B^k$  directed toward  $v_h^k$  and there is no positively weighted edge directed toward  $v_h^k$ , or else, among all the positively weighted edges directed toward  $v_h^k$ , the one in  $B^k$  has maximum weight.

Now, suppose that we have a  $y_h^{i+1}$  for each node  $v_h^{i+1}$  and a  $y_s^{i+1}$  for each set  $S$  of two or more nodes in  $G^{i+1}$ , such that (15) - (20) are satisfied (where  $B^0$  is replaced by  $B^{i+1}$ , etc.).

Let  $t_h^{i+1} = \sum y_s^{i+1}$ , summed over the sets  $S$  which contain node  $v_h^{i+1}$ .

To make the induction go through we assume further that in  $G^{i+1}$

(21) for every node  $v_h$ , such that  $t_h + y_h > 0$ , there exists at least one edge  $e_j$  directed toward  $v_h$  such that  $c_j = t_h + y_h$ .

This clearly holds for  $G^k$ , and we will prove from (15) - (21) for  $G^{i+1}$  that (15) - (21) holds for  $G^i$ .

Obtain the vector  $y^i$  as follows.

Where  $A$  is the set of nodes in circuit  $Q^i$  of  $G^i$ , where  $e_2^i$  is the edge of  $Q^i$  not in  $B^i$ , where  $v_2^i$  is the front end of  $e_2^i$ , where  $c_0^i$  is the minimum weight in  $Q^i$ , and where  $v_1^{i+1}$  is the node in  $G^{i+1}$  to which  $Q^i$  was shrunk, i.e.,

(22)  $y_2^i = y_1^{i+1} + c_2^i - c_0^i$ , and

$$(23) \quad y_A^1 = c_2^1 - y_2^1 - t_1^{i+1} .$$

Where  $v_3^1$  is any node in  $A$  other than  $v_2^1$ , and where  $e_4^1$  is the edge in  $Q^1$  which is directed toward  $v_3^1$ , let

$$(24) \quad y_3^1 = c_4^1 - y_A^1 - t_i^{i+1} .$$

Observe that (24) holds also for  $v_3^1 = v_2^1$ .

Where  $v_5^1$  is any node of  $G^1$  which is not in  $Q^1$ , let

$$(25) \quad y_5^1 = y_5^{i+1} .$$

Where  $R$  is a non empty subset of nodes in  $G^{i+1}$  which does not contain  $v_1^{i+1}$ , where  $J = R \cup v_1^{i+1}$ , where  $K = R \cup A$ , and where  $L$  is any set of two or more nodes in  $G^1$  such that  $L \cap A$  is a proper subset of  $A$ , let

$$(26) \quad y_R^1 = y_R^{i+1} ,$$

$$(27) \quad y_K^1 = y_J^{i+1} , \text{ and}$$

$$(28) \quad y_L^1 = 0 .$$

That completes the description of vector  $y^1$ . Now we must verify (15) - (21) for it .

For every edge of  $G^1$  which is directed toward a node not in  $A$ , for every node not in  $A$ , and for every set  $B$ , except  $A$ , in  $G^1$ , conditions (15) - (18), (20), and (21) follow immediately from those same conditions for  $y^{i+1}$ , (25) - (28), and the local nature of the change from  $G^{i+1}$ ,  $B^{i+1}$ , and  $c^{i+1}$  to  $G^1$ ,  $B^1$ , and  $c^1$ .

For every subset of nodes in  $G^1$  which does not contain all of  $A$ , condition (19) follows immediately as above. For set  $A$  and for every set  $K$  as in (27), condition (19) follows from (27), condition (19) for set  $J$  in  $G^{i+1}$ , and the fact that there are exactly  $|K| - |J| = |A| - 1$  more edges of  $B^1$  with both ends in  $K$  than there are edges of  $B^{i+1}$  with both ends in  $J$ , namely the edges of  $B^1 \cap Q^1$ .

It follows from (24), (27), and (28), that (21) holds for every node  $v_3^1$  in  $A$  (in particular where  $e_j$  is the  $e_4^1$  of (24)), and that (20) holds for every edge of  $B^1 \cap Q^1$ , and that (17) holds for  $e_2^1$ .

Condition (18) follows immediately for each node of  $A$  except  $v_2^1$  since there is an edge of  $B^1 \cap Q^1$  directed toward it. If there is an edge  $e_1^{i+1}$  in  $B^{i+1}$  which is directed toward  $v_1^{i+1}$ , then  $e_1^1$  is an edge of  $B^1$  which is directed toward  $v_2^1$ , and so in this case (18) follows for  $v_2^1$ . Otherwise, if there is no edge of  $B^{i+1}$  directed toward  $v_1^{i+1}$ , then by (18) for  $v_1^{i+1}$ ,  $y_1^{i+1} = 0$ . Also in this case, the  $c_2^1$  of (22) was chosen in the algorithm to be  $c_0^1$ . Therefore, if there is no edge of  $B^{i+1}$  directed toward  $v_1^{i+1}$ , then (22) is  $y_2^1 = 0$ , and so (18) follows for  $v_2^1$ .

For  $e_1^i$ , the only edge, if any, which is in  $B^i - Q^i$  and directed toward a node in  $A$ , we have  $c_1^{i+1} = c_1^i + c_0^i - c_2^i$  (from (1)), (22),  $y_1^{i+1} + w_1^{i+1} = c_1^{i+1}$  which is (20) for  $e_1^{i+1}$ , and  $w_1^i = w_1^{i+1}$  from (27) and (28). Combining these we get  $y_2^i + w_1^i = c_1^i$ , which is (20) for  $e_1^i$ .

Thus conditions (18), (19), (20), and (21) are now completely accounted for. Condition (17) for edges not in  $Q^i$  but directed toward nodes in  $A$ , condition (16) for  $y_A^i$ , and condition (15) for nodes in  $A$ , remain to be verified.

Let  $e_5^i$  be any edge of  $G^i$  which has both ends in  $A$ , and let  $v_3^i$  be its front end. To prove (17) for  $e_5^i$ , which is  $y_3^i + w_5^i \geq c_5^i$  where  $w_5^i = y_A^i + t_1^{i+1}$ , combine (24) and  $c_4^i \geq c_5^i$ .

Let  $e_3^i$  be any edge of  $G^i$  which has its front end  $v_3^i$  in  $A$  and its rear end not in  $A$ . To prove condition (17) for  $e_3^i$ , which is  $y_3^i + w_3^i \geq c_3^i$  where  $w_3^i = w_3^{i+1}$ , combine (24), (23), (22), (1), and (17) for  $e_3^{i+1}$ .

To prove (16) for  $A$ , that is  $y_A \geq 0$ , we use (21) for  $v_1^{i+1}$ . Assuming  $t_1^{i+1} + y_1^{i+1} > 0$ , let  $e_3^{i+1}$  be the  $e_j$  of that relation, let  $v_3^i$  be the front end of  $e_3^i$  in  $A$ , and let  $e_4^i$  be the edge of  $Q^i$  which is directed toward  $v_3^i$ . Here (21) is  $c_3^{i+1} = t_1^{i+1} + y_1^{i+1}$ . In this case, obtain  $y_A \geq 0$  by combining (23), (22), (21) for  $v_1^{i+1}$ , (1), and (4).

If there is no  $e_3^{i+1}$  directed toward  $v_1^{i+1}$  such that  $c_3^{i+1} = t_1^{i+1} + y_1^{i+1}$ , then  $t_1^{i+1} + y_1^{i+1} = 0$ , and all edges directed toward  $v_1^{i+1}$  have negative weight in  $G^{i+1}$ , so none of them are in  $B^{i+1}$ . Therefore since in this case the  $c_2^i$  of (22) was chosen to be  $c_0^i$ , (22) becomes  $y_2^i = 0$ , and (23) becomes  $y_A^i = c_0^i$ . By (2), we have  $y_A^i \geq 0$ .

Prove (15) for any node  $v_3^i$  in  $A$  by combining (24), (23), (22), (3) and  $y_1^{i+1} \geq 0$ .

That completes the proof of Theorems 1 and 2.

### §8

Notice from the proof that if every weight  $c_j, e_j \in G$ , is an integer, then the vector  $y^0$ , as well as vector  $x^0$ , is integer-valued. In particular, where every  $c_j = 1$ , vector  $y^0$  is 0,1-valued and  $\max(c, x) = \min(b, y)$  is a simple "Konig-type" theorem, analogous to the maximum-cardinality-matching duality theorem in [1].

The following two theorems can be proved by the methods used here.

**Theorem 3.** Where  $(L_A)$  is  $\sum x_j = n$ , summed over all edges  $e_j \in G$ , the vertices of the polyhedron given by  $(L_1), \wedge^{(L_2)} (L_3)$ , and  $(L_4)$  are precisely the vectors of the  $n$ -cardinality subsets of edges in  $G$  which comprise branchings. (In particular, where  $n$  is one less than the number of nodes in  $G$ , these branchings are the spanning arborescences of  $G$ ).

The present research began when A.J. Goldman asked for a description of "the convex hull of the spanning trees of a graph." Theorem 4 is proved in [3].

Theorem 4 . The vertices of the polyhedron  $P_G$  given by  $(L_1)$  and  $(L_3)$  are precisely the vectors of the subsets of edges in  $G$  which comprise forests. The vertices of the intersection of  $P_G$  with  $(L_4)$  are a subset of the vertices of  $P_G$  .

#### References

- [1] Jack Edmonds, Paths, trees, and flowers, Canadian J. Math., 17(1965), pp. 449-467.
- [2] \_\_\_\_\_, Maximum matching and a polyhedron with 0,1 - vertices, J. Res. National Bureau of Standards, 69B(1965), pp. 125-130.
- [3] \_\_\_\_\_, Matroids and the greedy algorithm, to appear.
- [4] R.E. Gomory, The traveling salesman problem, Proceedings of the IBM Scientific Computing Symposium on Combinatorial Problems, 1966, pp. 93-117.
- [5] J.B. Kruskal, On the shortest spanning subtree of a graph and the traveling salesman problem, Proc. Amer. Math. Soc., 7(1956), pp. 48-50.
- [6] P. Rosenstiehl, Deux algorithmes de l'arbre minimum, International Seminar on Graph Theory, Rome, July 1966.

15

This paper was to have appeared in the published proceedings of the International Seminar on Graph Theory and Its Applications, Rome, July 1966, sponsored by the International Computation Center. Various international failures of communication during the editorial process precluded it. I am sorry to have lost that opportunity to record my contribution to an outstanding symposium. I wish to acknowledge here my appreciation to the organizers of the symposium for the excellent job they did and for their kindness to me.

(

AN INTRODUCTION TO MATCHING

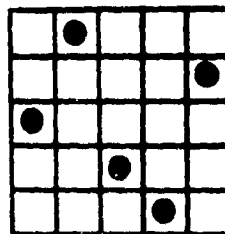
Jack Edmonds  
National Bureau of Standards  
Washington, D.C.

## I. The Optimum Assignment Problem

For rectangular array (matrix),  $N$ , we define a matching in  $N$  to be a subset  $M$  of the positions in  $N$  such that each column and each row of  $N$  contains at most one member of  $M$ .

For any square array  $N$ , we define a transversal or a perfect matching  $M$  to be a subset of the positions in  $N$  such that each column and each row of  $N$  contains exactly one member of  $M$ .

The optimum assignment problem is, given any  $n \times n$  array  $N$  of real numbers, find in  $N$  a transversal the sum of whose entries is maximum, i.e., an "optimum" transversal.



A transversal in  $N$  "assigns" the rows of  $N$  to the columns of  $N$ . Where the columns are people and the rows are jobs, and where each numerical entry represents the value of the person of that column at the job of that row, an optimum transversal represents an optimum assignment of the people to the jobs.

A well-known generalization of the assignment problem is the integer Hitchcock-transportation problem: Given a rectangular array  $N$  of real numbers,  $c_{ij}$ , and given an integer  $a_i \geq 0$  for each row  $i$  and an integer  $b_j \geq 0$  for each column  $j$ , assign a non-negative integer  $x_{ij}$  to each position  $(i,j)$  so that

$$(1) \text{ for every } i, \sum_j x_{ij} = a_i,$$

$$(2) \text{ for every } j, \sum_i x_{ij} = b_j,$$

and so that  $\sum_{i,j} c_{ij} x_{ij}$  is minimum (or maximum).

If  $a_i$  represents the number of refrigerators available at factory  $i$ , and  $b_j$  represents the number of refrigerators ordered by dealer  $j$ , and  $c_{ij}$  represents the cost of shipping a refrigerator from  $i$  to  $j$ , then  $\sum_{i,j} c_{ij} x_{ij}$  represents the total cost of the particular manner  $[x_{ij}]$  of distributing the refrigerators.

The assignment problem is where all  $a_i = 1$  and all  $b_j = 1$ .

A minor variation of the assignment problem is: given a rectangular array  $N$  of real numbers find in  $N$  a matching whose entries have maximum sum. This variation corresponds to replacing the equality signs in (1) and (2) by inequality signs. Ofcourse, a maximum matching will not contain a position whose entry is negative. In particular if all the entries are negative then the maximum matching will be the empty matching. It is an interesting exercise to discover how any maximum transversal problem can be solved by solving a maximum matching problem, and vice versa.

There are other generalizations and variations of the optimum assignment problem -- most notably integer network flow problems. These lectures, after treating the assignment problem itself will deal (briefly, I'm afraid) with some bizarre variations.

In an  $n \times n$  array there are  $n!$  different transversals. In particular there  $100!$  ways to assign 100 people to 100 jobs.  $100!$  is very large. If our method for finding an optimum assignment spent one microsecond per possible assignment, it would take hundreds of years to optimally assign 100 men.

It is a remarkable fact there exists a consistently good algorithm. An algorithm good enough that you could actually do as homework any instance of the assignment problem with 100 people, 100 jobs, and any collection of 3 digit numbers as values. Good enough to be used for many thousands of people and jobs. Ofcourse you have to know how, and it is not easy to discover how.

It is an unfortunate fact for most combinatorial problems -- problems very similar to the assignment problem -- that good algorithms are not known. For most such problems, though they are ofcourse finite, the best known methods do considerably worse than one might expect. Such problems include the bulk of integer linear programming problems.

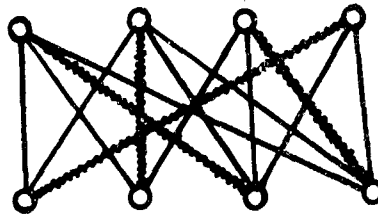
Therefore we do what we can. We experiment with the most promising methods we can find for problems that need answers. And we also try to find classes of problems for which, using special methods, we can predict computational efficiency. These lectures fall

into the latter area.

For ease in giving an arm-waving description of an algorithm, it is convenient to represent the assignment problem by a graph.

A bipartite graph  $G$  is one whose set  $V$  of nodes partitions into two sets  $V_1$  and  $V_2$  so that no edge of  $G$  has both ends in the same set. Thus, each edge meets one node in  $V_1$  and one node in  $V_2$ . Denote the set of edges of  $G$  as  $E$ .

A matching  $M$  in a graph is subset of its edges such that no two members meet the same node.



A perfect matching in a graph is a subset of its edges such that exactly one member meets each node.

The optimum assignment problem (more precisely, a minor generalization of it) is:

In any given bipartite  $G$ , with a real numerical weight  $c_e$  for each edge  $e \in E$ , find if there is one a perfect matching  $M$  which maximizes  $\sum_{e \in M} c_e$ , i.e. "a maximum perfect matching".

After we treat the above problem, we shall treat the same problem where  $G$  is not necessarily bipartite. The latter is a very substantial generalization.

Where  $G$  is bipartite, with "parts"  $V_1$  and  $V_2$ , the nodes of  $V_1$  correspond to rows or jobs the nodes of  $V_2$  correspond to columns or people. Each edge corresponds to a position in the array.

Any perfect matching in  $G$  determines an assignment of people to jobs, as does a perfect matching in the array.

Where  $G$  is any graph not necessarily bipartite, the maximum perfect matching problem is the problem of optimally pairing-off a set of objects (the nodes). Admissible pairs and their values are represented by the edges.

We shall see later that various other problems reduce to the matching problem.

A feasible node-weighting of graph  $G$  is a vector  $[y_v]$  with a component  $y_v$  for each node  $v \in V$  such that, for every edge  $e \in E$ ,

$$(3) \quad y_u + y_w \geq c_e \quad \text{where } u \text{ and } w \text{ are the ends of } e.$$

Lemma 1. For any perfect matching  $M$  of a graph  $G$  and for any feasible node-weighting  $[y_v]$ ,

$$(4) \quad \sum_{e \in M} c_e \leq \sum_{v \in V} y_v.$$

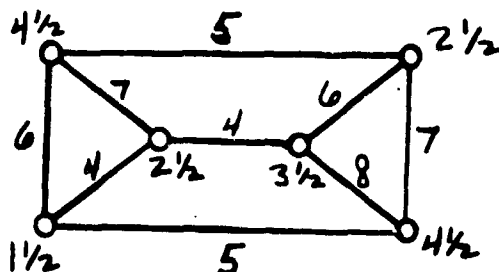
Proof: Add up the inequalities (3) which correspond to the edges in  $M$ .

It follows from Lemma 1 that if we can find a perfect matching  $M$  and a feasible  $[y_v]$  such that equality holds in (4), then

$\sum_{e \in M} c_e$  must be maximum and  $\sum_{v \in V} y_v$  must be minimum.

This is true whether or not  $G$  is bipartite. When one can get such a node-weighting along with an  $M$ , it provides a very simple guarantee that the  $M$  is maximum.

If  $G$  is not bipartite, there does not necessarily exist a feasible  $[y_v]$  such that  $\sum y_v$  equals the maximum value of  $\sum_{e \in M} c_e$  for  $G$ .



For any given edge-weighted bipartite graph  $G$ , the optimum assignment algorithm, that we will describe, first chooses any feasible node-weighting  $[y_v]$ . Then it chooses a matching, not necessarily perfect. Then it successively finds better node-weightings and better matchings until

- (1) it finds a node-weighting and a perfect matching for which

$$\sum_{e \in M} c_e = \sum_{v \in V} y_v, \text{ or until}$$

- (2) it finds a way to choose node-weightings such that  $\sum_{v \in V} y_v \rightarrow \infty$  (in which case, by Lemma 1, there is no perfect matching in  $G$ ).

Thus, the algorithm will prove

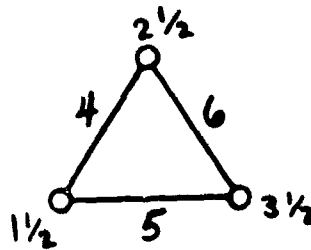
**Theorem 2.** For any edge-weighted bipartite graph  $G$ , which contains a perfect matching, the maximum weight-sum of a perfect matching in  $G$  equals the minimum sum of a feasible node-weighting.

P.S. 1 If  $G$  contains no perfect matching, then there is no minimum feasible node-weight sum.

By using only integer node-weights, if edge-weights are integers, the algorithm will also prove:

P.S. 2 If the edge-weights are all integers, then a min node-weighting can be chosen to be integers.

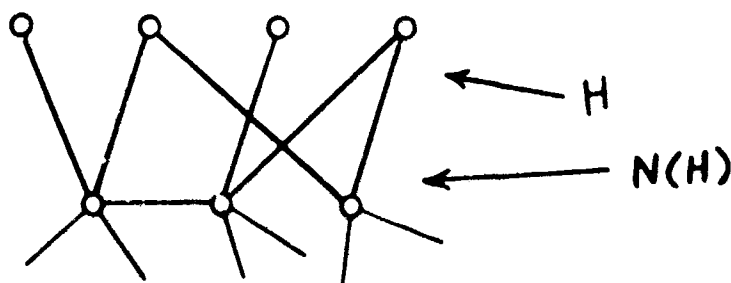
The two postscript as well as the theorem, are false for non-bipartite graphs.



## II. A Hungarian Method

We back up now and state Theorem 1, which will be used as part of the optimum assignment algorithm, to prove Theorem 2 already given.

A subset  $H$  of the nodes in a graph  $G$  is called hungarian, relative to  $G$ , if no two of them are joined by an edge of  $G$  and if the set  $N(H)$  of neighbors of  $H$  has fewer members than  $H$ , i.e.,  $|N(H)| < |H|$ . A neighbor of  $H$  is a node which is joined by an edge of  $G$  to a node in  $H$ .



Obviously, a graph which contains a hungarian set  $H$  can not contain a perfect matching  $M$ , because the  $|H|$  members of  $M$  which meet the nodes in  $H$  would have to meet  $|H|$  different members of  $N(H)$ . This is impossible since  $|N(H)| < |H|$ .

A non-bipartite graph does not necessarily contain either a hungarian set or a perfect matching.



Theorem 1. A bipartite graph  $G$  contains no perfect matching if and only if  $G$  has a hungarian set, contained either in part  $V_1$  or part  $V_2$  of  $G$ .

Subroutine R1 of the algorithm will prove Thm 1 by finding in any bipartite graph either a perfect matching or a hungarian set. Subroutine R2 of the algorithm will prove Thm 2 by using a hungarian set of a subgraph of  $G$  to improve any non-minimum node-weighting of  $G$ .

Here is the algorithm, Given a bipartite graph  $G$  with a numerical weight  $c_e$  on each edge  $e \in E$ .

Give to  $G$  any feasible node-weighting  $[y_v]$ . (Make the node-weights integers if the edge-weights are.)

For the current feasible node-weighting at any stage of the algorithm, let  $G'$  denote the subgraph of  $G$  which consists of all nodes of  $G$  and those edges  $e$  of  $G$  such that

$$(5) \quad y_u + y_v = c_e, \text{ where } u \text{ and } v \text{ are the ends of } e.$$

We call  $G'$  the equality subgraph of  $G$  relative to  $[y_v]$ .

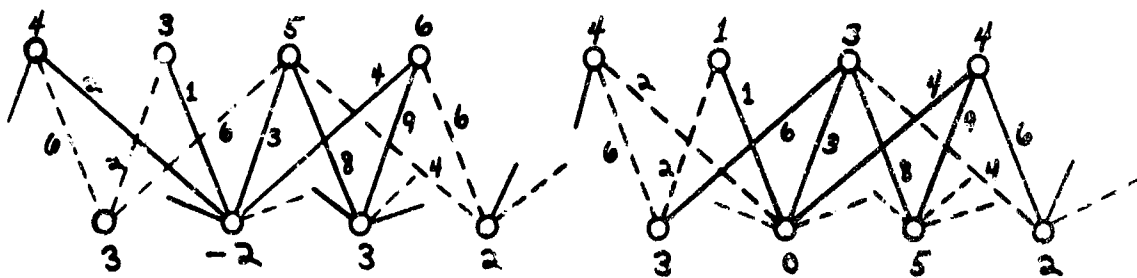
Using routine R1, we find either a perfect matching  $M$  in  $G'$  or else a hungarian set  $H'$  of  $G'$  in either part  $V_1$  or  $V_2$ . (Since every edge of  $G$  has one of its end-nodes in set  $V_1$  and its other end-node in set  $V_2$ , the same is true for subgraph  $G'$ . Any perfect matching of  $G'$  is ofcourse also a perfect matching of  $G$ , since  $G'$  contains all the nodes of  $G$ . However, a hungarian set  $H'$  of  $G'$  is not in general a hungarian set of  $G$ . Since  $G$  generally has more edges than  $G'$ , set  $H'$  of nodes generally has more neighbors relative to  $G$  than relative to  $G'$ .)

Theorem 1 says that  $G'$  has either an  $M$  or an  $H'$ . Routine R1 will be the proof of Theorem 1.

Suppose we find a perfect matching  $M$  in  $G'$ . Then adding together equations (5),  $y_u + y_w = c_e$ , for the edges  $e$  in  $M$ , we get  $\sum_{e \in M} c_e = \sum_{v \in V} y_v$ , and so  $M$  is optimum.

Otherwise, we find a hungarian set  $H'$  of  $G'$  contained in, say, part  $V_2$ . In this case, we apply routine R2:

R2. Suppose that the set  $N(H')$ , of neighbors of  $H'$  relative to  $G'$ , is not the entire set of neighbors of  $H'$  relative to  $G$ . Then there are edges  $e$  of  $G$  which are not in  $G'$  and which have one end in  $H'$  and the other end not in  $N(H')$ . Let  $\epsilon = \min (y_u + y_w - c_e)$  over all such edges. Clearly,  $\epsilon > 0$ . Lowering each node-weight in  $H'$  by  $\epsilon$  and raising each node-weight in  $N(H')$  by  $\epsilon$ , we get a new feasible node-weighting. Its sum is smaller because  $|N(H')| < |H'|$ . The equality subgraph  $G'$  changes, so we apply R1 again.



Where  $N(H')$  is the entire set of neighbors of  $H'$ , relative to  $G$  as well as  $G'$ , the set  $H'$  is hungarian relative to  $G$  and so there is no perfect matching in  $G$ . In this case, we can take  $\epsilon$

to be as large as we please, and still change node-weights as above.

This gives  $\sum v$  as small as we please, i.e.  $\sum v \rightarrow -\infty$ .

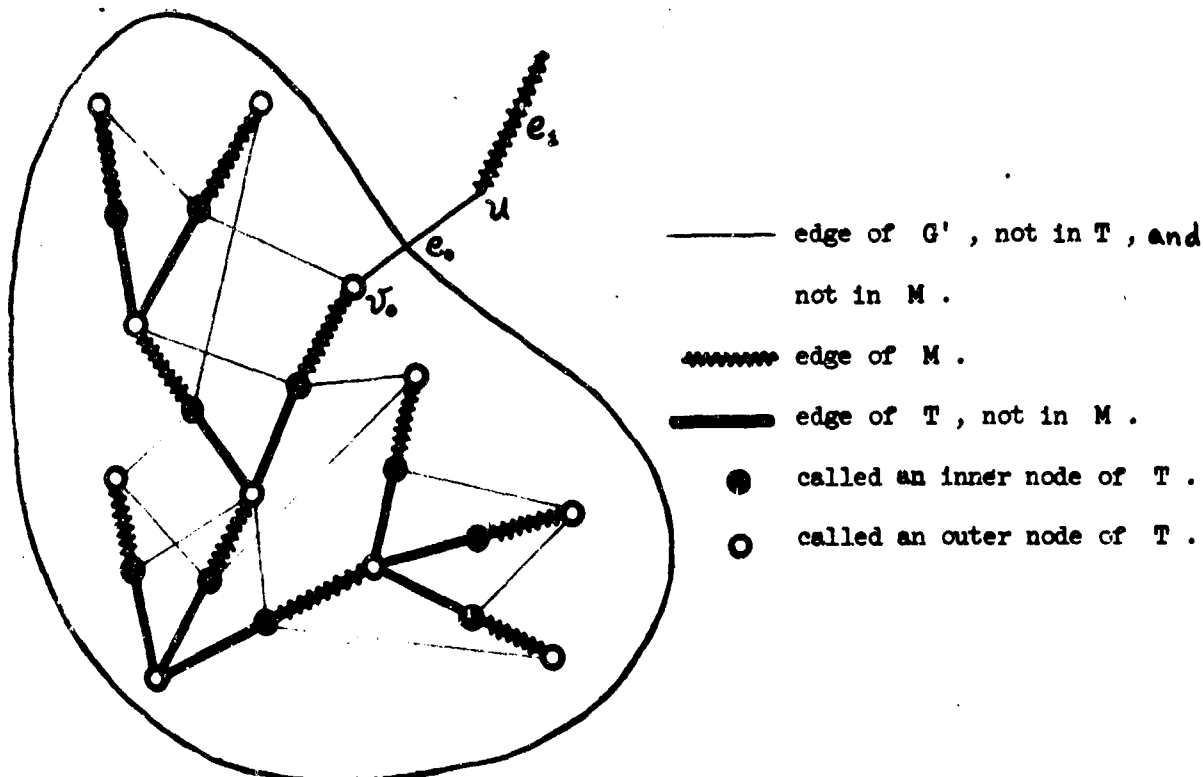
Assuming Thm 1 holds, i.e., assuming there is a valid routine R1, the routine R2 just described proves Thm 2 and PS1. For a node-weighting having min sum, there must be a perfect matching in the corresponding  $G'$ , because otherwise we can apply Thm 1 and R2 to get a smaller node-weight sum.

PS2 asserted that if edge-weights are integers then a min node-weighting can be chosen to be integers. This follows by applying the above process to any integer-valued feasible node-weighting, because whenever edge-weights and node-weights are integers,  $\epsilon$  is an integer (or arbitrary  $-\infty$ ).

Let us now describe routine R1, thereby proving Thm 1 and completing the description of an optimum-assignment algorithm. R1 must find either a perfect matching in  $G'$  or a hungarian set of  $G'$ .

R1: Let  $M$  be any matching in  $G'$ , not necessarily a perfect matching. To begin it might be the empty matching. If  $M$  is not perfect, let  $r$  be any node which  $M$  does not meet. We shall find either a hungarian set of  $G'$  which contains  $r$ , or else a matching  $M'$  in  $G'$  which is better than  $M$  in the sense that  $M'$  meets  $r$  as well as all the nodes which  $M$  meets.

In  $G'$ , "grow a tree"  $T$  of the kind pictured inside the bean. Node  $r$  itself is a tree of type  $T$ . Start off with simply it.



All the outer nodes of  $T$  must be in the same part of  $G'$ . Thus, no two of them are joined by a edge of  $G'$ . The number of inner nodes of  $T$  is exactly one less than the number of outer nodes.

For any  $T$  either (a), (b), or (c) holds. (a) Every edge in  $G'$  which meets an outer node of  $T$  meets an inner node of  $T$  at its other end. In this case, the set of outer nodes is a hungarian set  $H$  of  $G'$ . The set of inner nodes is  $N(H)$ , relative to  $G'$ .

Otherwise some edge  $e_0$  of  $G'$  meets an outer node, say  $v_0$ , of  $T$  and a node, say  $u$ , not in  $T$ . (b)  $u$  meets no edge in  $M$ . In this case obtain  $M'$  from  $M$  by interchanging the matching roles of edges in the path  $P$ , consisting of  $u$ ,  $e_0$ , and the path from  $v_0$  to  $r$  in  $T$ . Then forget  $T$  and, if there is still some node, say  $r'$ , in  $G'$  which does not meet any edge of  $M'$ , grow in  $G'$  another tree rooted at  $r'$ . (c) Otherwise,  $u$  meets an edge  $e_1 \in M$ . In this case, enlarge  $T$  by adjoining  $e_0$  and  $e_1$  to it, so that  $u$  becomes an inner node and the other end of  $e_1$  becomes an outer node.

Suppose the number of nodes in  $G$ , and thus in  $G'$ , is  $2n$ . After  $T$  is enlarged at most  $n$  times, either (a) or (b) must occur. After at most  $n$  differently-rooted trees are grown, either (a) or a perfect matching must occur in  $G'$ .

When (a) occurs, we apply R2 to the hungarian set  $H$  of  $G'$ , the set of outer nodes in  $T$ . Thus, unless  $H$  is also a hungarian set relative to  $G$ , the node weights change and the equality subgraph  $G'$  changes. The outer nodes of  $T$  are then not a hungarian set of the new  $G'$ . However, the same matching  $M$  is contained in the new  $G'$  and the same tree  $T$  is contained in the new  $G'$ . Case (b) or (c) of R1 applies directly to this  $M$  and  $T$  in the new  $G'$ , and so we can continue growing the same tree in the new  $G'$ . Thus, R1 is iterated a total of at most  $n^2$  times and the "step" of enlarging a tree is iterated a total of at most  $n^2$  times before one obtains either a perfect matching in some equality subgraph  $G'$  or else a hungarian set of  $G$ .

### III. Bipartite Matching and Linear Programs

By polyhedron (strictly speaking convex polyhedron) we mean the set of all vectors (points) which satisfy some given finite collection of linear equations and linear inequalities.

Let  $G$  be any finite graph, not necessarily bipartite. Let  $E$  denote the set of edges of graph  $G$ . Let  $V$  denote the set of nodes of  $G$ .

Let there be a real variable  $x_e$  for each edge  $e \in E$ . Let  $P_G$  be the polyhedron of vectors  $[x_e]$  such that

(1) for every  $e \in E$ ,  $x_e \geq 0$ , and

(2) for every  $v \in V$ ,  $\sum x_e = 1$ , where the sum is taken over all edges  $e$  which meet node  $v$ .

Another way of expressing (1) and (2) is

(1')  $x_j \geq 0$ ,  $j \in E$ , and

(2')  $\sum_j a_{ij} x_j = b_i$ ,  $j \in E$ ,  $i \in V$ ,

where all  $b_i = 1$ , where  $a_{ij} = 1$  if edge  $j$  meets node  $i$ , and where  $a_{ij} = 0$  if edge  $j$  does not meet node  $i$ .

Matrix  $[a_{ij}]$  is called the incidence matrix of graph  $G$ . It has a "row"  $i$  for each node  $i \in V$  and a "column"  $j$  for each edge  $j \in E$ . Clearly, a matrix is the incidence matrix of some graph if and only if each of its columns contains exactly two 1's and the rest 0's.

We may define a vertex  $x^0$  of a polyhedron  $P$  to be a point (i.e., a vector  $[x_j^0] = x^0$ ) in  $P$  such that, for some linear function  $U = \sum_j c_j x_j$  of the points in  $P$ ,  $x^0$  is the only point in  $P$  which maximizes the function.

It is a standard theorem that if a given linear function of points in polyhedron  $P$  has a maximum (i.e.,  $P$  is not empty and the function is not unbounded above), then the function is maximized by a vertex of  $P$ . Ofcourse, some linear functions on  $P$  are maximized by other points as well.

As you know, the linear programming problem is: For the polyhedron  $P$ , determined by a given system of linear "constraints", and for a given linear function  $U$  of points in  $P$ , find a vertex of  $P$  which maximizes (or minimizes)  $U$  in  $P$ .

Any vector  $[x_e]$  of zeroes and ones is called the incidence vector of (or simply the vector of) the subset of  $e$ 's such that  $x_e = 1$ .

Thus, every subset of edges in  $G$  is represented by a unique 0,1 vector, and conversely.

Clearly, where  $c_e$  is the weight on edge  $e$  in  $G$ , the weight-sum of any perfect matching  $M$  in  $G$  is the value of  $\sum_e c_e x_e$  ( $e \in E$ ) for the vector of  $M$ .

Clearly, the 0,1-valued vectors contained in  $P_G$  (in fact, the integer-valued vectors contained in  $P_G$ ) are precisely the vectors of perfect matchings in  $G$ .

We shall show that the assignment problem is an instance of linear programming by showing that

Theorem 3. If  $G$  is bipartite, then the vertices of  $P_G$  are precisely the vectors of perfect matchings in  $G$ .

Clearly, even if  $G$  is not bipartite, the vector of any perfect matching  $M$  in  $G$  is a vertex of  $P_G$ , since we can display a function  $\sum_e c_e x_e$  which obviously is maximized in  $P_G$  only by the vector of  $M$ . In particular, where  $c_e = 1$  if and only if  $e \in M$ .

The hard part is to show that every vertex of  $P_G$  is the vector of a matching. We shall do so using the duality thm of linear programming and thm 2. about node-weightings for an edge-weighted bipartite  $G$ .

The l.p. dual of maximizing  $U = \sum_j c_j x_j$ , subject to  $x_j \geq 0$  and  $\sum_i a_{ij} x_j = b_i$ , is minimizing  $W = \sum_i b_i y_i$  subject to  $\sum_j a_{ij} y_i \leq c_j$ . The duality thm says that  $\max U = \min W$  if these extrema exist.

In particular, the dual of maximizing  $U = \sum_e c_e x_e$  ( $e \in E$ ) in  $P_G$  is minimizing  $W = \sum_v (v \in V)$  subject to  $y_u + y_w \geq c_e$  for every  $e$ , where  $u$  and  $w$  are the end-nodes of  $e$ . That is, minimizing the sum of feasible node-weights, as in Thm 2.

Thus, it follows from Thm 2 and the l.p. duality thm that if  $G$  is bipartite, then, for any  $U$ , the max of  $U$  for vectors in  $P_G$  equals the max of  $U$  for vectors of perfect matchings.

Therefore, since all vectors of perfect matchings are in  $P_G$ ,  $U$  is maximized in  $P_G$  by the vector of a perfect matching.

For any vertex  $x^0$  of  $P_G$ , suppose that  $U$  is a linear function that is maximized in  $P_G$  only at  $x^0$ . Then  $x^0$  must be the vector of a perfect matching. So Thm 3 is proved.

Conversely, Thm 3 and the l.p. duality thm immediately imply Thm 2. So, in view of l.p. duality, Thm 2 and Thm 3 are equivalent.

Where bipartite graph  $G$  is a square array whose rows  $i$  and columns  $j$  are the nodes, and whose positions  $(i,j)$  are the edges, Thm 3 is well-known as G. Birkhoff's theorem on "doubly stochastic matrices" (1946)

An  $n$  by  $n$  doubly stochastic matrix is defined to be an  $n$  by  $n$  matrix  $[x_{ij}]$  such that:

- all  $x_{ij} \geq 0$ , and
- (3) for every fixed  $i$ ,  $\sum_j x_{ij} = 1$ , and
- (4) for every fixed  $j$ ,  $\sum_i x_{ij} = 1$ . (see page )

Matrices are vectors, indexed differently. The collection of  $n$  by  $n$  doubly stochastic matrices is a polyhedron. The Birkhoff thm says that the vertices of this polyhedron are the  $n$  by  $n$  permutation matrices. A permutation matrix is a matrix such that there is a 1 in each row, a 1 in each column, and all other entries are zeroes.

Thm 2, essentially, is due to Egervary (1931). The algorithm here for the assignment problem, essentially, is due to Kuhn and Munkres (1955-57).

Where the 1 in (3) is replaced by any prescribed integers  $a_i \geq 0$  and the 1 in (4) is replaced by any prescribed integers  $b_j \geq 0$ , we get the linear constraints of the integer transportation problem, relations (1) and (2) of section I. These constraints, together with  $x_{ij} \geq 0$  for every  $i$  and  $j$ , define a polyhedron, say  $P_T$ .

Theorem 3 readily generalizes to the fact that all the vertices of  $P_T$  are integer-valued vectors. Thus we have the very well-known fact that the integer transportation problem is an instance of linear programming.

Indeed, historically the first, and still the most prominent, algorithms for the transportation problem are direct applications of the simplex method.

Theorem 3, and consequently Theorem 2, and consequently even Theorem 1, are readily (and often) proved using general l.p. techniques together with the special properties of the incidence matrix  $[a_{ij}]$  of a bipartite graph.

Similarly, extremal (integer) flows in a network can be treated by applying general l.p. techniques to matrices  $[a_{ij}]$  which are the "incidence matrices of directed graphs". And, on the other hand, the combinatorial algorithm that we described for the assignment problem is very closely related to the combinatorial methods of Ford and Fulkerson for network flow problems.

Two advantages of these combinatorial methods are (1) for practical purposes, they are more efficient than simplex methods, and (2) for theoretical purposes, they provide theoretical bounds on

efficiency of computation that are so far not available for simplex methods. The disadvantage of these combinatorial methods is that they have not been satisfactorially extended to general l.p. problems.

#### IV. Matching in a general graph

So far these lectures have been small variations on old stuff. Now I would like to show one of the ways in which these combinatorial methods can be generalized to certain problems which can not be treated directly as linear programs -- at least not in the usual sense. Actually they are linear programs determined implicitly by astronomical collections of linear constraints. Surprisingly, these problems are just as tractable as the assignment problem -- in spite of the traditional views as to why the assignment problem is so tractable.

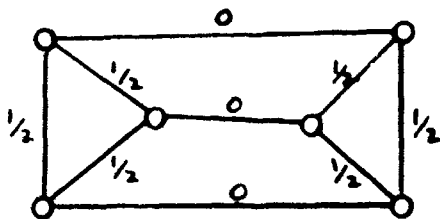
We treat the problem of finding a maximum-weight perfect matching in an edge-weighted graph  $G$  which is not necessarily bipartite. In particular we can take  $G$  to be a complete graph -- a graph in which every pair of nodes is joined by an edge.

Recall that we have already defined a polyhedron  $P_G$ , for any graph  $G$ , by the linear constraints (1) and (2) in section III. We have already observed that the only integer-valued vectors contained in  $P_G$  are the vectors of perfect matchings in  $G$ . That is, an integer-valued vector  $[x_e]$  is contained in  $P_G$  if and only if its components are 0's and 1's, and the 1-components correspond to the edges of a perfect matching.

We have already observed that where  $c_e$  is the weight on  $e$  in  $G$ , the weight-sum of any perfect matching  $M$  is the value of  $\sum c_e x_e (e \in E)$  for the vector of  $M$ .

Hence, the maximum perfect-matching problem, for any edge-weighted graph  $G$  is an instance of "integer linear programming". Integer linear programming is the problem of maximizing a given linear function by an integer-valued vector subject to some given linear constraints.

If all the vertices of polyhedron  $P_G$  were vectors of matchings, as in the case where  $G$  is bipartite, then the maximum matching problem would be simply an instance of linear programming. However, when  $G$  is not bipartite,  $P_G$  has vertices which are fractional-valued.



The picture shows one of the simplest graphs  $G$  such that  $P_G$  contains a fractional vertex as well as several vertices which are vectors of perfect matchings. The numbers on the edges are the components of the fractional vertex

in this  $P_G$ . Notice that where the edge-weights are the numbers on the edges in the picture on page 6, the maximum-weight of a perfect matching is 18. At the fractional vertex of  $P_G$ ,  $\sum_{e \in E} c_e x_e$  has the value 19, which by no accident is also the min sum of a feasible node-weighting.

A main idea in tackling the general perfect-matching problem is to chop off the fractional vertices of  $P_G$  so as to obtain a polyhedron  $P'_G$  such that all the vectors of perfect matchings are still contained in  $P'_G$  and such that  $P'_G$  doesn't have any fractional vertices, i.e., all its vertices are vectors of perfect matchings. In other words, obtain the convex hull,  $P'_G$ , of the vectors of perfect matchings.

Theorem 4. For any graph  $G$ , the convex hull of the vectors of perfect matchings in  $G$  is the polyhedron  $P'_G$  given by the following linear constraints:

- (1) for every edge  $e \in E$ :  $x_e \geq 0$ ;
- (2) for every node  $v \in V$ :  $\sum x_e = 1$ , where the sum is taken over all edges  $e$  which meet node  $v$ .
- (3) for every subset  $s$  of nodes which has cardinality  $|s| = 2q_s + 1$  for some positive integer  $q_s$ ,  $\sum x_e \leq q_s$ , where the sum is taken over all edges  $e$  which have both ends in  $s$ .

Any vector of a matching, say  $M$ , in  $G$  satisfies (3) for any set  $s$ , since no more than  $q_s$  edges of  $M$  can have both ends in  $s$ , and since such edges are the only ones in  $M$  which appear in (3). Therefore, every vector of a perfect matching of  $G$  is in the polyhedron  $P'_G$ .

It is not so obvious that every vertex of  $P'_G$  is a vector of a perfect matching. We prove this by means of an algorithm -- in a manner analogous to our proof of Thm 3 by means of the assignment algorithm.

The idea of treating a combinatorial problem by chopping away at a polyhedron to eliminate undesirable vertices is by no means new. A long time ago Kuhn and Dantzig and others took this approach to the traveling salesman problem. I believe Motzkin tried it for the "3-dimensional assignment problem". Gomory discovered finite algorithms for integer linear programming which operate by chopping locally until the answer is a vertex of the resulting polyhedron. His methods could be adapted to give finite algorithms for describing the convex hull of the integer vectors in any given bounded polyhedron.

For the matching problem, and certain cousins, the chopping-idea has been completely successful in the following senses. (1) We get a succinct and useful description of precisely the relevant polyhedra. (2) We get an algorithm that is really good.

Later, I'll describe a representative, "optimum branching", of one other essentially different class of problems for which polyhedron-chopping has been completely successful.

Let  $U = \sum c_e x_e (e \in E)$  be any linear function of vectors  $[x_e]$ , determined by an arbitrary specification of edge-weights for  $G$ . To maximize  $U$  by a vertex of  $P'_G$  is a linear program. Our purpose is to solve this l.p. by the vector of a perfect matching. Since  $U$  is arbitrary and since for every vertex of  $P'_G$  there is a  $U$  which is maximized only by that vertex, this will prove Thm 4.

We now describe the dual of our l.p. Like l.p. duals everywhere, it has a variable for each constraint of the primal, other than non-negativity, and it has a constraint for every variable of the primal, as well as the non-negative constraint on each dual variable that corresponds to an inequality-constraint of the primal.

In particular, it has a variable  $y_v$  for each node  $v \in V$ , the "node-weight" for  $v$ . And it has variable  $y_s$ , an "odd-set-weight", for each subset  $s \subset V$  such that  $|s| = 2q_s + 1$  where  $q_s$  is a positive integer. Recall that  $|s|$  means the number of nodes in  $s$ .

The variables  $y_v$  are allowed to go negative since they correspond to equations of the primal. The constraints of the dual are

(4) for every set  $s$ ,  $y_s \geq 0$ ;

(5) for every edge  $e$ ,  $f_e(y) = y_u + y_w + \sum y_s \geq c_e$ ,

where  $u$  and  $w$  are the ends of  $e$ , and where the summation is over all sets  $s$  which contain both ends of  $e$ .

Thus, a dual weighting  $y = [y_v, y_s]$  is called feasible if (4) and (5) hold.

The linear function to be minimized is

(6)  $W = \sum y_v + \sum q_s y_s$ .

The duality theorem tells us that a vector  $x^0 = [x_e^0]$  maximizes  $U = \sum c_e x_e$  subject to (1), (2), and (3) if and only if there is some vector  $y^0 = [y_v^0, y_s^0]$  satisfying (4) and (5), for which  $U(x^0) = W(y^0)$ . We shall find such an  $x^0$  which is the vector of a perfect matching.

More directly useful to our purpose is the so-called "complementary slackness theorem" on dual l.p.'s. For our particular dual l.p.'s it says that a vector  $x^0 = [x_e^0]$  maximizes  $U$  subject to (1), (2), and (3), and a vector  $y^0 = [y_v^0, y_s^0]$  minimizes  $W$  subject to (4) and (5), if and only if

(7) for every variable  $y_s$ , either  $y_s^0 = 0$  or else equality holds in the corresponding constraint (3);

(8) for every variable  $x_e$ , either  $x_e^0 = 0$  or else equality holds in the corresponding constraint (5).

Where  $x^0 = [x_e^0]$  is the vector of a perfect matching  $M$  of  $G$ ,

(7) says that for every set  $s$  of nodes, either  $y_s^0 = 0$  or else  $q_s$  edges of  $M$  have both ends in  $s$ ;

(8) says that for every edge  $e$  in  $M$ , equality holds in the corresponding constraint (5).

The matching algorithm finds a perfect matching  $M$  and a vector  $y^0 = [y_v^0, y_s^0]$  such that (4), (5), (7), and (8) hold. This guarantees that  $M$  is optimum and it also proves Thm 4.

Vectors  $[y_v, y_s]$  have an awful lot of components. Fortunately, however, the ones we deal with have no more non-zero components than the number of edges in  $G$ .

## V. A Matching Algorithm

Given any edge-weighted graph  $G$ . The algorithm starts out just like for the bipartite case. We choose a feasible node-weighting  $[y_v]$ , i.e., values of  $y_v$  such that (5) holds with all  $y_s = 0$ . Let  $G'$  be the equality subgraph of  $G$ , relative to this  $[y_v, y_s]$ . That is,  $G'$  consists of all nodes of  $G$  and those edges  $e$  for which equality holds in the corresponding constraint (5),  $f_e = c_e$ .

If we can find a perfect matching in  $G'$  then it will be optimum because then it and the dual weights will satisfy (4), (5), (7), and (8). Generally we will not be able to find a perfect matching in this  $G'$  or in any other such  $G'$  determined by a node-weighting. However, lets try to, just as in the bipartite case.

Choose any matching  $M$ , not necessarily perfect, in  $G'$ . If there is a node  $r$  which  $M$  doesn't meet, start growing in  $G'$  a tree  $T$  rooted at  $r$ , just as we do for the bipartite case.

Recall that when  $G$  is bipartite, either (a), (b), or (c) must hold for  $T$  in  $G'$ . (See Section II).

Bipartite or not, when we spot an occurrence of (c), we enlarge  $T$  in  $G'$ . Bipartite or not, when we spot an occurrence of (b), we get a better matching in  $G'$ , we discard the current  $T$  in  $G'$ , and if the matching is still not perfect we start another tree  $T$ .

In case (a), the outer nodes of  $T$  comprise a hungarian set  $H$  relative to  $G'$ , and the inner nodes of  $T$  comprise the set  $N(H)$  of neighbors of  $H$  relative to  $G'$ . Bipartite or not, when (a) occurs we change the dual weighting  $[y_v, y_s]$  so that other edges of  $G$  enter  $G'$ . We then continue to treat the same  $T$  relative to the new  $G'$ . In general, the way the dual-weighting changes is more complicated than for a bipartite  $G$  where we don't have any positive weights  $y_s$ . Indeed, so far our weights  $y_s$  are all zero. We shall have to describe how some of them become positive. We'll do so after we describe the concept of pseudo-node.

When  $G$  is not bipartite, a fourth case (d) can occur:

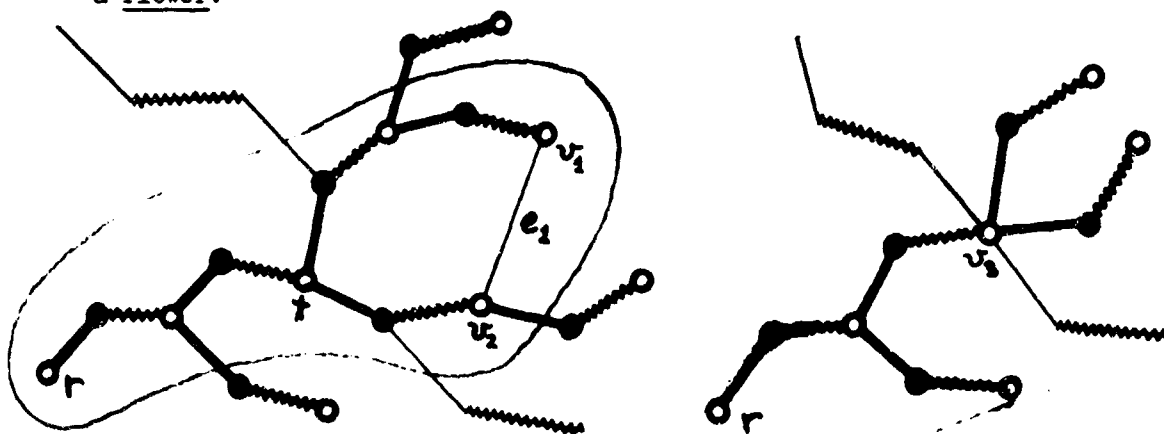
(d) Two outer nodes, say  $v_1$  and  $v_2$ , of  $T$  are joined by an edge, say  $e_1$ , of  $G'$ .

When  $G$  is bipartite, any two outer nodes of  $T$  must be in the same part,  $V_1$  or  $V_2$ , of  $G$ , and so (d) can not occur.

One can immediately verify that for any  $G'$  (regardless of whether  $G$  is bipartite), either (a), (b), (c), or (d) holds for  $T$  in  $G'$ .

When (d) occurs, matters get especially tricky. .

Let  $P_1$  be the path in  $T$  from  $v_1$  back to  $r$ . Let  $P_2$  be the path in  $T$  from  $v_2$  back to  $r$ . Paths  $P_1$  and  $P_2$  together with the edge  $e_1$  joining  $v_1$  and  $v_2$  form what we call a flower.



It consists of a stem: the path between  $r$  and  $t$ ; and a blossom  $B$ : the polygon. Ofcourse if  $P_1$  and  $P_2$  happen not to run together until they get back to  $r$ , then  $t = r$  and hence the stem is just the node  $r$ . It is also possible for  $t = v_1$ , or even  $t = r = v_1$ . The number of edges and the number of nodes in blossom  $B$  is odd and greater than 1, i.e.,  $2q + 1$  where  $q$  is some positive integer.

We now obtain a new graph  $G$ , a new subgraph  $G'$ , and a new tree  $T$ , from the ones we've got, by shrinking to a single pseudo-node,  $v_3$ , the blossom  $B$  and all edges that have both ends in  $B$ .

The edges of the matching  $M$  that are not shrunk away do form a new matching  $M$  in the new  $G'$ . The new  $T$ , formed by the edges of the old  $T$  which are not shrunk away, is a tree in the new  $G'$  having the correct structure relative to the new  $M$ . The pseudo-node  $v_3$  is an outer-node of the new  $T$ . If  $t = r$ , then  $v_3$  is the root of the new  $T$ .

Whenever we shrink a blossom we remember it so that later we can expand the pseudo-node to recover it. We never bother to remember the edges of the current matching in a blossom that we shrink, because they are not likely to be compatible with the matching that is current when we expand the pseudo-node back to the blossom.

The algorithm continues as before, considering occurrences of (a), (b), (c), or (d), relative to the new  $G$ ,  $G'$ ,  $M$ , and  $T$ .

Each time we spot an occurrence of (d), we shrink the blossom, thereby obtaining still another  $G$ ,  $G'$ ,  $M$ , and  $T$ . A blossom containing pseudo-nodes might be shrunk into another pseudo-node, so we can have pseudo-nodes "inside" the pseudo-nodes of  $G'$ . The set, say  $s$ , of all real nodes inside a pseudo-node always has odd cardinality, because the sum of an odd number of odd numbers is odd. We remember every blossom we shrink so that any pseudo-node can be expanded anytime that it is not inside another pseudo-node.

Whenever we spot an occurrence of (c), we enlarge the  $T$  in the current  $G'$ .

Whenever we spot an occurrence of (b), we get a new matching in the current  $G'$  such that fewer nodes in  $G'$  are left unmatched by the new matching. When this happens we discard  $T$ . If there is another node  $r$  in  $G'$  still not met by the matching, we start growing in the same  $G'$  another tree rooted at that  $r$ . Pseudo-nodes of  $G'$  formed from blossoms of earlier trees may become inner nodes or outer nodes of this tree.

Whenever there are no cases of (b), (c), or (d) to spot, we have case (a), i.e., the outer nodes of  $T$  are a hungarian set  $H$  relative to  $G'$  and the inner nodes of  $T$  are the neighbor set  $N(H)$  of  $H$  relative to  $G'$ . We must in this case consider changing the dual weighting  $y = [y_v, y_s]$ .

In general, for the current feasible dual-weighting,  $y = [y_v, y_s]$ , a  $y_s$  is positive only if  $s$  is the set of real nodes inside some pseudo-node, either a pseudo-node of the current  $G'$  or a pseudo-node at any level inside a pseudo-node of the current  $G'$ . For every edge  $e$  which is either an edge of  $G'$  or an edge of a currently shrunk blossom, we have equality,  $f_e(y) = c_e$ , for the constraint (5) corresponding to  $e$ ; conversely, if  $f_e(y) = c_e$  holds for an edge  $e$  of the current  $G$ , i.e., for an edge  $e$  that is not inside a current pseudo-node, then  $e$  is an edge of  $G'$ . These are the senses in which a pseudo  $G'$  is the equality subgraph of a pseudo  $G$ .

Assuming these conditions hold, we now describe how, when (a) holds, to get a new feasible dual-weighting such that these conditions continue to hold and such that either we have a new  $Q'$  relative to which (b), (c), or (d) holds for  $T$ , or else we dispose of a pseudo inner node of  $T$ , or else we have  $W \rightarrow -\infty$ , in which case there is no perfect matching in  $G$ . We choose  $\epsilon$  to be as large as possible subject to the following constraints with right-hand sides given by the current dual-weighting.

(9) For every edge  $e$  of  $G$ , not in  $G'$ , such that one end of  $e$  is an outer node of  $T$  and the other end of  $e$  is not in  $T$ ,  
 $\epsilon \leq f_e(y) - c_e$ .

(10) For every edge  $e$  of  $G$ , not in  $G'$ , such that both ends of  $e$  are outer nodes of  $T$ ,  $2\epsilon \leq f_e(y) - c_e$ .

(11) For every  $s$  which is the set of all real nodes inside a pseudo inner node of  $T$ ,  $2\epsilon \leq y_s$ .

(The pseudo inner nodes that we refer to here are nodes of the current  $G$ , not pseudo nodes inside of pseudo nodes.)

We assume for the moment that at least one such constraint exists, so that  $\epsilon$  has a maximum.

Now we change the dual weighting  $[y_v, y_s]$  as follows. For every real node  $v$  which is either an outer node of  $T$  or else inside a pseudo outer node of  $T$ , lower  $y_v$  by  $\epsilon$ . For every  $s$

which is the set of all real nodes inside a pseudo outer node of  $T$ , raise  $y_s$  by  $2\epsilon$ . For every real node  $v$  which is either an inner node of  $T$  or else inside an inner node of  $T$ , raise  $y_v$  by  $\epsilon$ . For every  $s$  which is the set of all real nodes inside a pseudo inner vertex of  $T$ , lower  $y_s$  by  $2\epsilon$ .

Suppose that the size of  $\epsilon$  was determined by equality in some instance of either (9) or (10). Let  $e$  denote the corresponding edge. After the  $\epsilon$ -adjustment of the dual-weighting, the new  $G'$  is a certain different subgraph of the same  $G$  (perhaps having pseudo-nodes). Subgraph  $G'$  is determined by  $G$  and the new dual-weighting. The same  $T$  and  $M$  are in this new  $G'$ . The edge  $e$  enters  $G'$ . If only one end of  $e$  is an outer node of  $T$  (and the other end not in  $T$ ), then we have immediately an occurrence of either (b) or (c). If both ends of  $e$  are outer nodes of  $T$ , then we immediately have an occurrence of (d), a blossom to shrink as previously described.

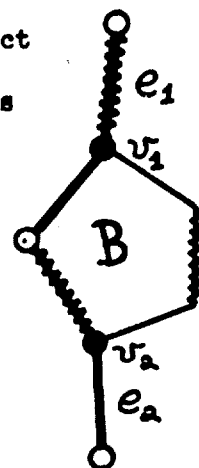
Suppose the size of  $\epsilon$  was determined by equality in some instance of (11). Let  $v$  denote the corresponding pseudo inner node. Let  $e_1$  denote the edge of the current matching which meets  $v$ . Edge  $e_1$  is in  $T$ . Let  $e_2$  denote the other edge of  $T$  which meets  $v$ . After the  $\epsilon$ -adjustment of the dual-weighting, we expand  $v$  to the blossom, say  $B$ , whose shrinking introduced  $v$ . This expansion gives rise to a new  $G$  and  $G'$ . It is easy to verify that  $B$  is part of the new  $G'$ . That is,  $f_e(y) = c_e$  holds for the edges  $e$  of  $B$ . Let  $v_1$  denote the node of  $B$  which edge  $e_1$  meets;

let  $v_2$  denote the node of  $B$  which  $e_2$  meets. The only node of  $B$  which is met by the current matching is  $v_1$ ; we are able to add to the matching certain edges of  $B$  so as to get a new matching  $M$ , in the new  $G'$ , which meets all the nodes of  $B$ .

The edges of the current  $T$  remain in the new  $G'$ . Unless

$v_1 = v_2$ , they do not form a tree in the new  $G'$ . However, they, *(a certain one of the two paths in  $B$  joining  $v_2$  to  $v_1$ )* together with  ~~$B$ , minus the non-matching edge in  $B$  which meets~~

~~$v_2$~~ , do form our new tree  $T$ . This new  $T$  does have the correct structure relative to the new  $G'$  and new  $M$ . Relative to this new  $T$ ,  $M$ ,  $G'$ ,  $G$ , and  $[y_v, y_g]$ , we now return to looking for further occurrences of (a), (b), (c), or (d). Incidentally, in the situation we just treated, i.e., where the size of  $\epsilon$  was determined by equality in an instance of (11), it might be that  $\epsilon = 0$ . This isn't relevant to the treatment.



We have finished describing all situations of the algorithm, except the two terminal situations.

One of the terminal situations is when we obtain in some  $G'$ , perhaps with pseudo nodes, a matching  $M$  which is perfect. We then expand pseudo nodes  $v$ , one after another, to the polygons  $B$  that they represent. Immediately before a pseudo node  $v$  is expanded, the matching  $M$ , that we have, is perfect in the graph  $G'$  with with node  $v$ , that we have. Let  $e$  denote the unique edge of  $M$  which, in that  $G'$ , meets node  $v$ . Expanding  $v$  to polygon  $B$  gives us a larger  $G'$  containing  $B$  instead of  $v$ . In this  $G'$ ,

$e$  is the only edge of  $M$  that meets a node of  $B$ . By adjoining to  $M$  certain edges of  $B$ , we obtain a perfect matching  $M$  for this larger  $G'$ . Unless there are no more pseudo-nodes, we then treat some pseudo-node of this  $G'$ , perhaps one in  $B$ , in the same way.

Eventually, we get a perfect matching  $M$  in the original graph  $G$ . It will be an optimum perfect matching, because it and the dual-weighting, that we have, together satisfy conditions (4), (5), (7), and (8).

The other terminal situation is an occurrence of (a) for which there are no constraints (9), (10), or (11). In this case,  $\epsilon$  can be chosen as large as we please. By using  $\epsilon \rightarrow \infty$  to change the dual weighting  $[y_v, y_g]$ , as already described, we get a feasible  $[y_v, y_g]$  such that  $W \rightarrow -\infty$ . Hence, there can be no perfect matching in  $G$ .

This completes the description of the algorithm. Just as when  $G$  is bipartite, we can observe a bound, relative to the size of  $G$ , on the number of operations in applying the algorithm to  $G$ , which shows the algorithm to be not only finite, but very good. At the same time, Theorem 4 is proved.

## VI. Theorems of Tutte, Peterson, and Konig

The algorithm also provides proof of the following theorem of W.T. Tutte (1947), analogous to Theorem 1. We define a Tutte family in a graph  $G$  to be a family of disjoint connected subgraphs  $G_i$  of  $G$  such that each  $G_i$  contains an odd number of nodes (perhaps one node) and such that, upon shrinking every  $G_i$  to a node  $v_i$ , the set of nodes  $v_i$  is a hungarian set of the resulting graph. (Tutte does not describe the family in this way.)

Theorem 5. A graph contains a perfect matching if and only if it does not contain a Tutte family.

The "only if" part is fairly easy. If  $G$  contains a Tutte family of subgraphs  $G_i$  and also a perfect matching  $M$ , then, because  $G_i$  has an odd number of nodes, at least one edge say  $e_i$  of  $M$  has one end in  $G_i$  and the other end not in  $G_i$ . After shrinking, the edges  $e_i$  meet the nodes  $v_i$ , and have distinct nodes at their other ends. This is impossible, however, since the set of nodes  $v_i$  is hungarian.

To prove the "if" part recall the terminal situation of the algorithm where, for an occurrence of (a), there are no constraints (9), (10), and (11). The absence of a constraint (11), means that no inner node of  $T$  is pseudo. Since the outer nodes of  $T$  are a hungarian set  $H$  of  $G'$ , the absence of constraints (9) and (10) means that  $H$  is also a hungarian set of  $G$ . Its neighbor set

$N(H)$  consists of the inner nodes of  $T$ , all real. The nodes in  $H$ , call them nodes  $v_1$ , are obtained by shrinking disjoint connected subgraphs  $G_1$  of the original, real-noded,  $G$ . Each  $G_1$  contains an odd number of nodes. Therefore, the original, real-noded,  $G$  contains a Tutte family.

The only alternative is the other terminal situation of the algorithm, and it yields a perfect matching. Thus, Tutte's theorem is proved.

Ofcourse, a much more pleasant proof can be obtained by stripping-down the algorithm to one for simply finding in any  $G$  either a perfect matching or a Tutte family. Indeed, finding an algorithm for the latter was a main hurdle in finding the algorithm that maximizes weight-sum. Tutte's original proof of Theorem 5 is fascinatingly unalgorithmic, and it prompted a number of programmatic efforts on the subject.

The subject of matchings started over 75 years ago with the 4-color map conjecture. The conjecture, still unproved, says that for any way of dividing up the plane into a "map", by a connected graph (a "planar" one) embedded in the plane so that every edge lies on the boundary of two different regions, the regions can be colored with only four colors so that any two regions having an edge in common are colored differently. The property of every edge lying on the boundary of two different regions of the map is equivalent to the planar graph containing no isthmus. An isthmus

of a connected graph, planar or not, is an edge whose deletion leaves the graph unconnected.

By "perturbing" at each node the conjecture easily reduces to the case where the graph has degree 3 at each node. The degree of a graph  $G$  at a node is the number of edge-ends in  $G$  that meet the node.

An interesting theorem, which we won't prove, is that: A planar map having degree 3 at every node can be colored (properly) with 4 colors if and only if the graph of the map contains three mutually disjoint perfect matchings. Thus, the 4-color conjecture is equivalent to the statement that any 3-degree, connected, planar graph with no isthmus contains 3 mutually disjoint perfect matchings.

In 1891, Peterson made the following contribution toward proving the 4-color conjecture.

Theorem 6. A 3-degree, connected graph  $G$  with no isthmus, whether planar or not, contains a perfect matching.

Let's prove this using Tutte's theorem. Suppose a graph  $G$  as described in Theorem 6 contains no perfect matching. Then it contains a Tutte family of subgraphs  $G_1$ . Since  $G_1$  contains an odd number of nodes, since each node has odd degree, and since the collection of edges meeting nodes in  $G_1$  has collectively an even number of edge-ends, the number of edges of  $G$  having one end in  $G_1$ , and one end not in  $G_1$ , is odd.

Not every  $G_1$  can have as many as 3 such edges, since the shrunken  $G_1$ 's are a hungarian set whose smaller neighbor set must meet all such edges, and no node in the neighbor set can meet more than 3 such edges. Therefore, there is at least one subgraph  $G_1$  such that exactly one edge, say  $e$ , has one end in  $G_1$  and one end not in  $G_1$ . Edge  $e$  is then an isthmus of  $G$ , contradicting the hypothesis. So Theorem 6 is proved.

By deleting the edges of a perfect matching  $M$  from the  $G$  of Theorem 6, we are left with simply a collection of mutually disjoint polygons. Clearly, the set of edges in this collection of polygons can be partitioned into two perfect matchings if and only if each of the polygons contains an even number of edges. Considerable effort has been spent on trying to prove that when  $G$  is planar, there exists an  $M$  such that  $G-M$  consists of even polygons.

If the  $G$  of Theorem 6 is bipartite, then, for any  $M$ ,  $G-M$  consists of even polygons, because it is easy to show that every polygon in a bipartite graph is even. Thus, the 4-color conjecture is proved for any planar map whose graph  $G$  is 3-degree and bipartite.

Indeed, though little is known about partitioning the edges of a general  $k$ -degree graph into  $k$  perfect matchings, we do have the following theorem about  $k$ -degree bipartite graphs. The elementary theory of bipartite graphs, including this theorem and Theorem 1, is due to Konig (circa 1925).

Theorem 7. The edges of any  $k$ -degree bipartite graph can be partitioned into  $k$  perfect matchings.

For any perfect matching  $M$  in a  $k$ -degree bipartite graph  $G$ , clearly  $G-M$  is a  $(k-1)$  degree bipartite graph. Hence, it suffices to show that any  $k$ -degree bipartite  $G$  contains an  $M$ . If  $G$  doesn't contain an  $M$ , then, by Theorem 1, it contains a hungarian set  $H$ . Together the  $|H|$  nodes in  $H$  meet  $k \cdot |H|$  different edges. At the other ends of all these edges is  $N(H)$ . But this is impossible, since the degree of each node in  $N(H)$  is only  $k$ , and  $N(H) < |H|$ . So Theorem 7 is proved.

The mystery of the 4-color conjecture seems not due to mystery about planar graphs and maps. The subject of "planarity" is very well understood. The mystery is due to the lack of a satisfactory theory about the combinatorics of coloring, i.e., partitioning. If you could find a good algorithm for deciding, for any given 3-degree graph  $G$ , whether the edges of  $G$  can be partitioned into 3 perfect matchings, then you could probably settle the 4-color conjecture.

## VII. Degree-Constrained Subgraphs

Given any graph  $G$  with a real numerical weight  $c_e$  for each edge  $e \in E$  and an integer  $b_v$  for each node  $v \in V$ , find in  $G$ , if there is one, a subgraph  $M$  which has degrees  $b_v$  at nodes  $v$  and whose edges have maximum weight-sum. This is called the "optimum  $b$ -matching problem" or the "optimum degree-constrained subgraph problem" (for "undirected graphs").

Where  $b = [b_v]$ ,  $v \in V$ , a  $b$ -matching  $M$  in  $G$  is a subset  $M \subset E$  of edges such that  $b_v$  edge-ends of edges in  $M$  meet node  $v$ . (Tutte and many other authors in graph theory would say " $b$ -factor" rather than " $b$ -matching".) Obviously there is a 1-1 correspondence between the  $b$ -matchings in a graph  $G$  and the  $b$ -degree subgraphs of  $G$  that contain all the nodes of  $G$ .

We allow  $G$ , and hence  $M$ , to contain loops and multiple parallel edges. A loop is an edge such that both of its ends meet the same node. Several edges are said to be parallel to each other if they all meet the same one or two nodes. Loops and multiple parallel edges are superfluous in the 1-matching problem.

Tutte (1954) generalized his Theorem 5 of the last section to a characterization of graphs  $G$  which, for given  $b = [b_v]$ , do not contain a  $b$ -matching. Using similar devices we shall show how to get a good algorithm for optimum  $b$ -matching

In fact, we shall generalize further, thereby including directly the integer network flow of problem of Ford and Fulkerson. The latter may be regarded as the following: Given any directed graph (network)  $G$  with a real numerical weight (cost)  $c_e$  for each edge (arc)  $e \in E$  and with an integer  $b_v$  for each node  $v \in V$ , find in  $G$ , if there is one, a subset  $M \subset E$  of edges such that, for every  $v \in V$ , the number of edge-ends of  $M$  directed toward  $v$  minus the number of edge-ends of  $M$  directed away from  $v$  equals  $b_v$ , and such that  $\sum_{e \in M} c_e$  is maximum (or minimum). A negative  $b_v$  is called a source, and a positive  $b_v$  is called a sink.

To get the appropriate common generalization of undirected graph and directed graph, we introduce the concept of "bidirected graph". A graph  $G$  is called bidirected if each edge-end of  $G$  has either a +1 or a -1 on it. Equivalently, each edge-end is directed either toward or away from the node it meets, independently of how the other end of the same edge is directed. Equivalently, each end of an edge is either a front-end or a rear end, independently of what the other end of the same edge is.

The degree of a node  $v$  in a bidirected graph  $G$  is the number of front-ends in  $G$  that meet  $v$  minus the number of rear ends in  $G$  that meet  $v$ . With this new definition of degree, the optimum degree-constrained subgraph problem is the same as stated above.

A bidirected  $G$  is directed if every edge of  $G$  has a front end and a rear end. A bidirected  $G$  is undirected if every edge of  $G$  has two front ends. Another interesting case is where every

edge has either two front ends or two rear ends.

Another generalization is obtained by introducing single ended objects called slacks, positive slacks and negative slacks, i.e., front slacks and rear slacks. A slack in a graph meets only one node. A graph  $G$  with slacks is regarded as undirected when all ends in  $G$  including those of slacks are front ends.

Slacks conveniently represent upper and lower bound degree-constraints. Suppose we wish to find a maximum weight subgraph  $M$  of  $G$  such that the degree of  $M$  at node  $v$  is at least  $b_v^1$  and at most  $b_v^2$ . In other words, suppose we wish to find in  $G$  a maximum weight  $b$ -matching where  $b = [b_v]$  and where each  $b_v$  is in the interval  $b_v^1 \leq b_v \leq b_v^2$ . Obtain from  $G$  a new graph  $G_0$  by introducing at each node  $v$  of  $G$ ,  $b_v^2 - b_v^0$  negative slacks and  $b_v^0 - b_v^1$  positive slacks where  $b_v^0$  is some integer between  $b_v^1$  and  $b_v^2$ . Give these slacks weight zero. Finding a maximum weight  $b_v^0$ -matching in  $G_0$  is equivalent to finding a maximum weight  $b$ -matching in  $G$ .

You may ask why I don't introduce "edges" with 3 ends. I would if I knew a good algorithm for handling them.

Another valuable generalization, suggested by the transportation and integer flow problems, is to maximize  $U = \sum c_e x_e (e \in E)$  by an integer-valued vector  $x = [x_e]$  that satisfies

- (1) for every element  $e \in E$ , either an edge  $e$  or a slack  $e$ ,

$$0 \leq x_e \leq a_e;$$

(2) for every node  $v \in V$ ,

$$\sum_e a_{ev} x_e = b_v, (e \in E), \text{ where}$$

$a_{ev} = 0$  if  $e$  does not meet  $v$ ,

$a_{ev} = 1$  if  $e$  has one front end at  $v$ ,

$a_{ev} = -1$  if  $e$  has one rear end at  $v$ ,

$a_{ev} = 2$  if  $e$  is a loop with two front ends at  $v$ , and

$a_{ev} = -2$  if  $e$  is a loop with two rear ends at  $v$ .

Matrix  $[a_{ev}]$  is called the incidence matrix of bidirected graph

$G$ . Set  $V$  is the set of nodes in  $G$ . Set  $E$  is the set of edges, including loops, and also the slacks in  $G$ . The integer  $b_v$  is the degree-constraint at  $v$ . The integer  $\alpha_e$  is called the capacity on  $e$ .

When every  $\alpha_e$  is 1, this problem is simply the  $b$ -matching problem relative to graph  $G$ . When the  $\alpha_e$ 's are any positive integers, the problem is the  $b$ -matching problem relative to the graph obtained from  $G$  by replicating  $\alpha_e$  times the element  $e$ . We also permit  $\alpha_e = \infty$ . Of course, if the problem has a solution it must be possible to replace  $\alpha_e = \infty$  by a large  $\alpha_e$ . However, as a matter of fact, infinite capacities are much easier to handle than finite capacities.

We now describe how any optimum  $b$ -matching problem can be reduced to an optimum 1-matching problem.

We describe first how any  $b$ -matching problem on a bidirected graph  $G$  can be reduced to a  $b^*$ -matching problem on an undirected graph  $G^*$ . For each node  $v$  in  $G$ , let there be two nodes, say  $u$  and  $w$ , in  $G^*$ . Let all the front ends at  $v$  be front ends at  $u$  in  $G^*$ . Let all the rear ends at  $v$  be front ends at  $w$  in  $G^*$ . Let there be in  $G^*$  a new edge  $e$  with a front end at  $u$ , a front end at  $w$ , and any

appropriately large capacity. Let the degree constraints  $b_u^*$  and  $b_v^*$  be appropriately large and such that  $b_u^* - b_v^* = b_e$ . Let every edge or slack in  $G$  have the same weight in  $G^*$  as in  $G$ . Let the new edges in  $G^*$  have weight zero.

One can verify that if  $G$  is directed then  $G^*$  is bipartite.

We next describe how any  $b$ -matching problem on an undirected graph  $G$  can be reduced to a  $b^*$ -matching problem on an undirected graph  $G^*$  such that the edge capacities are all  $\infty$  and such that there are no slacks.

For each edge  $e$  of  $G$  having finite capacity  $\alpha_e$ , and meeting say nodes  $u$  and  $v$  in  $G$ , replace  $e$  by a path  $P_e$  of three edges joining  $u$  to  $w$ ; give the two new nodes, interior to  $P_e$ , degree constraints equal to  $\alpha_e$ . Give one of the non-middle edges of  $P_e$  the weight that  $e$  had; give zero weights to the other two edges of  $P_e$ . Let  $r$  denote a special new node. For each slack  $e$  of  $G$ , having capacity  $\alpha_e$ , and meeting say node  $v$  in  $G$ , replace  $e$  by a path  $P_e$  of two edges joining  $v$  to  $r$ ; give the new node, interior to path  $P_e$ , a degree-constraint equal to  $\alpha_e$  or  $b_v$ , whichever is smaller. Give the edge of  $P_e$  that meets  $v$  the weight that slack  $e$  had; give zero weight to the other edge of  $P_e$ . Let there be a zero-weighted loop with both (front) ends at  $r$ . Give  $r$  any appropriately large degree-constraint whose parity is such that the sum of the degree-constraints on all nodes is even. The result of this construction is the desired  $b^*$ -matching problem.

Next we describe how to reduce any  $b$ -matching problem on <sup>an undirected</sup> graph  $G$ , such that there are no slacks and such every edge  $e$  has capacity  $\alpha_e = \infty$ , to a perfect matching problem (like treated in section V) on

a graph, say  $G^*$ . For every node  $v$  in  $G$ , having degree-constraint  $b_v$ , make  $b_v$  copies of  $v$  to be nodes of  $G^*$ . Join a pair of nodes of  $G^*$  by an edge, say  $e^*$ , if and only if these two nodes are copies of nodes in  $G$  that are joined by an edge, say  $e$ , of  $G$ . Let the weight of the  $e^*$  in  $G^*$  equal the weight of the  $e$  in  $G$ . Let every node in  $G^*$  have degree-constraint equal 1. The pre-image in  $G$ , with appropriate multiplicities, of an optimum perfect matching in  $G^*$ , is an answer to the given  $b$ -matching problem.

To use these reductions directly as algorithms is computationally rather wasteful. However, one can use them to derive a good direct algorithm for  $b$ -matching from the algorithm for 1-matching.

A paper to appear, called "Optimum degree-constrained subgraphs," by Ellis Johnson and me, will describe a direct algorithm, theorems analogous to Theorem 4, and computational experience, for general optimum  $b$ -matchings.

The matching algorithm described in section V was devised in 1962. Since then I've tried to find good algorithms for some other obviously finite problems. I have found a few, but it seems that such algorithms are not easy to come by.

# Minimum Partition of a Matroid Into Independent Subsets<sup>1</sup>

Jack Edmonds

(December 1, 1964)

A matroid  $M$  is a finite set  $M$  of elements with a family of subsets, called independent, such that (1) every subset of an independent set is independent, and (2) for every subset  $A$  of  $M$ , all maximal independent subsets of  $A$  have the same cardinality, called the rank  $r(A)$  of  $A$ . It is proved that a matroid can be partitioned into as few as  $k$  sets, each independent, if and only if every subset  $A$  has cardinality at most  $k \cdot r(A)$ .

## 1.0. Introduction

Matroids can be regarded as a certain abstraction of matrices [8].<sup>2</sup> They represent the properties of matrices which are invariant under elementary row operations but which are not invariant under elementary column operations—namely properties of dependence among the columns. For any matrix over any field, there is a matroid whose elements correspond to the columns of the matrix and whose independent sets of elements correspond to the linearly independent sets of columns. A matroid  $M$  is completely determined by its elements and its independent sets of elements.

The same letter will be used to denote a matroid and its set of elements. The letter  $I$  with various sub or superscripts will be used to denote an independent set.

The interest of matroids does not lie only in how they generalize some known theorems of linear algebra. There are examples, which I shall report elsewhere, of matroids which do not arise from any matrix over any field—so matroid theory does truly generalize an aspect of matrices. However, matroid theory is justified by new problems in matrix theory itself—in fact by problems in the special matrix theory of graphs (networks). It happens that an axiomatic matroid setting is most natural for viewing these problems and that matrix machinery is clumsy and superfluous for viewing them. The situation is somewhat similar to the superfluity of (real) matrices to the theory of linear operators, though there a quite different aspect of matrices is superfluous. When it comes to implementing either theory, matrices are often the way to do it.

Matroid theory so far has been motivated mainly by graphs, a special class of matrices. A graph  $G$  may be regarded as a matrix  $N(G)$  of zeroes and ones, mod 2,

which has exactly two ones in each column. The columns are the edges of the graph and the rows are the nodes of the graph. An edge and a node are said to meet if there is a one located in that column and that row. Of course a graph can also be regarded visually as a geometric network. It is often helpful to visualize statements on matroids for the case of graphs, though it can be misleading. Matroids do not contain objects corresponding to nodes or rows.

Theorem 1 on "minimum partitions," the subject of this paper, was discovered in the process of unifying results described in the next paper, "On Lehman's Switching Game and a Theorem of Tutte and Nash-Williams" (denoted here as "Part II"), which is a direct sequel. Theorem 1 is shown there to be closely related to those results. Lately, I have learned that Theorem 1 for the case of graphs (see sec. 1.7) was anticipated by Nash-Williams [5].

By borrowing from work of others, I intend that this paper together with possible sequels be partly expository and technically almost self-contained.

## 1.1. The Problem

Various aspects of matroids—in particular, the first pair of axioms we cite—hold intrinsic interest which is quite separate from linear algebra.

**AXIOM 1:** Every subset of an independent set of elements is independent.

Any finite collection of elements and family of so-called independent sets of these elements which satisfies axiom 1 we shall call an independence system. This also happens to be the definition of an abstract simplicial complex, though the topology of complexes will not concern us.

It is easy to describe implicitly large independence systems which are apparently very unwieldy to analyze. First example: given a graph  $G$ , define an independent set of nodes in  $G$  to be such that no edge of  $G$  meets two nodes of the set. Second example: define an independent set of edges in  $G$  to be such that

<sup>1</sup>Sponsored by the Army Research Office (Durham). Presented at the Seminar on Matroids, National Bureau of Standards, Aug. 31-Sept. 11, 1964. I am much indebted to Alfred Lehman for encouraging my interest in the subject.

<sup>2</sup>Figures in brackets indicate the references at the end of this paper.

no node meets two edges of the set. Third example: define an independent set of edges in  $G$  to be such that the edges of the set, as column vectors of  $N(G)$ , are linearly independent. The third example is the prototype of the systems we shall study here.

A *minimum coloring* of the nodes of a graph  $G$  is a partition of the nodes into as few sets (colors) as possible so that each set is independent. A good characterization of the minimum colorings of the nodes in a graph is unknown (unless the graph is bipartite, i.e., the nodes can be colored with two colors). To find one would undoubtedly settle the "four color" conjecture.

A problem closely related to minimum coloring is the "packing problem." That is to find a good characterization (and an algorithm) for maximum cardinality independent sets. More generally the "weighted packing problem" is, where each element of the system carries a real numerical weight, to characterize the independent sets whose weight-sums are maximum. The packing problem for the systems of the first example is also very much unsolved (unless the graph is bipartite).

The minimum coloring problem for the systems of the second example is unsolved (unless the graph is bipartite). Its solution would also undoubtedly settle the four-color conjecture. However the packing problem, and more generally the weighted packing problem, is solved for the second example by the extensive theory of "matchings in graphs."

For the third example the packing problem is in a sense trivial. It is well known that the system of linearly independent sets of edges in a graph, and more generally the system of linearly independent sets of columns in a matrix, satisfies the following:

**AXIOM 2:** For any subset  $A$  of the elements, all maximal independent sets contained in  $A$  contain the same number of elements.

A *matroid* is a (finite) system of elements and sets of elements which satisfies axioms 1 and 2.

For any independence system, any *subsystem* consisting of a subset  $A$  of the elements and all of the independent sets contained in  $A$  is an independence system. Thus, a matroid is an independence system where the packing problem is postulated to be trivial for the system and all of its subsystems. For me, having spent much labor on packing problems, it is pleasant to study such systems. Matroids have a surprising richness of structure, as even the special case of graphic matroids shows.

Clearly, a subsystem of a matroid  $M$  is a matroid. We call it a *submatroid* and we use the same symbol to denote it and its set of elements. The *rank*,  $r(A)$ , of a set  $A$  of elements in  $M$  or the *rank*,  $r(A)$ , of the submatroid  $A$  of  $M$  is the number of elements in each maximal independent set contained in  $A$ , i.e., the number of elements in a base of  $A$ .

The main result of this paper is a solution of the minimum coloring problem for the independent sets of a matroid. Another paper will treat the weighted packing problem for matroids.

## 1.2. Ground Rules

One is tempted to surmise that a minimum coloring can be effected for a system by some simple process like extracting a maximal independent set to take on the first color, then extracting a maximal independent set of what is left to take on the second color, and so on till all elements are colored. This is usually far from being successful even for matroids, though it is precisely matroids for which a similar sort of monotonic procedure always yields a maximum cardinality independent set and, as we shall see, in another paper, also always yields a maximum weight-sum independent set when the elements carry arbitrary real weights.

Consider the class of matroids implicit in the class  $M_F$  of all matrices over fields of integers modulo primes. (For large enough prime, this class includes the matroid of any matrix over the rational field.) We seek a good algorithm for partitioning the columns (elements of the matroid) of any one of the matrices (matroids) into as few sets as possible so that each set is independent. Of course, by carrying out the monotonic coloring procedure described above in all possible ways for a given matrix, one can be assured of encountering such a partition for the matrix, but this would entail a horrendous amount of work. We seek an algorithm for which the work involved increases only algebraically with the size of the matrix to which it is applied, where we regard the size of a matrix as increasing only linearly with the number of columns, the number of rows, and the characteristic of the field. As in most combinatorial problems, finding a naive algorithm is trivial but finding an algorithm which meets this condition for practical feasibility is not trivial.

We seek a good characterization of the minimum number of independent sets into which the columns of a matrix of  $M_F$  can be partitioned. As the criterion of "good" for the characterization we apply the "principle of the absolute supervisor." The good characterization will describe certain information about the matrix which the supervisor can require his assistant to search out along with a minimum partition and which the supervisor can then use with ease to verify with mathematical certainty that the partition is indeed minimum. Having a good characterization does not mean necessarily that there is a good algorithm. The assistant might have to kill himself with work to find the information and the partition.

Theorem 1 on partitioning matroids provides the good characterization in the case of matrices of  $M_F$ . The proof of the theorem yields a good algorithm in the case of matrices of  $M_F$ . (We will not elaborate on how.) The theorem and the proof apply as well to all matroids via the matroid axioms. However, the "goodness" for matrices depends on being able to carry out constructively with ease those matrix operations which correspond to the existential assertions of the theory. A fundamental problem of matroid theory is to find a good representation for general matroids—good perhaps relative to the rank and the number of elements in the matroids. There is a very

elegant lattice representation (geometric lattices, [1, 2]), but it is not something you would want to record except for the very simplest matroids.

### 1.3. The Theorem

The cardinality of a set  $A$  is denoted by  $|A|$ . The rank of a set  $A$  is denoted by  $r(A)$ .

**THEOREM 1:** *The elements of a matroid  $M$  can be partitioned into as few as  $k$  sets, each of which is independent, if and only if there is no subset  $A$  of elements of  $M$  for which*

$$|A| > k \cdot r(A).$$

The theorem makes sense for any independence system  $M$  if we define the rank  $r(A)$  of any subset  $A$  to be the maximum cardinality of an independent set in  $A$ . In fact, the "only if" part of the theorem is true for any independence system  $M$ . Let  $I_i (i=1, \dots, k)$  be  $k$  independent sets in  $M$  for which

$$\bigcup_{i=1}^k I_i = M.$$

For any subset  $A$  of  $M$ ,  $|I_i \cap A| \leq r(A)$  and

$$|A| \leq \sum_{i=1}^k |I_i \cap A| \leq k \cdot r(A).$$

Thus the "only if" part is proved.

In general for the coloring problem in nonmatroidal systems, the other half of the theorem is not true. However, the Konig theorem on matchings in bipartite graphs can be regarded as a valid instance of theorem 1 for certain nonmatroidal systems. A bipartite graph is a graph whose nodes can be partitioned into two sets each independent (by coincidence, an instance of the coloring problem in our first example). The Konig theorem says that for a bipartite graph  $G$  the minimum number of nodes which meet all the edges equals the maximum number of edges such that no node meets more than one of them. (This theorem solves the packing problem for a special case of our second example of independence system.)

Fourth example: For a graph  $G$ , let the elements of the system  $M$  be the edges of  $G$ . For each node of  $G$ , let the set of edges which meet the node be an independent set in  $M$ . Let the subsets of these sets be the rest of the independent sets in  $M$ . The Konig relation for a graph  $G$  implies theorem 1 for system  $M$ .

Theorem 1 for the system  $M$  arising from  $G$  does not imply the Konig theorem for  $G$ . For independence systems in general the relation represented by theorem 1 is weaker than the relation represented by the Konig theorem—the latter being that the minimum number of independent sets which together contain all the elements equals the maximum number of elements in a set of rank one. It's nice to have the weaker relation of theorem 1 because it might apply to other systems where the well known Konig relation does not.

### 1.4. Terminology

There are various families, (1) through (6), of subsets of the elements in a matroid  $M$  which are used in describing the structure of  $M$ .

(1) The family of independent sets of  $M$ .

(2) The family of minimal dependent sets of elements in  $M$  (where dependent means not independent). These are called the *circuits* in  $M$ . The letter  $C$  with various sub or superscripts will be used to denote a circuit.

(3) The family of *spans* or *closed sets* in  $M$ . A *span*  $S$  in  $M$  is a set of elements such that no circuit of  $M$  contains exactly one element not in  $S$ . That is,  $|S \cap C| \neq 1$  for every circuit  $C$  in  $M$ .

The *span* or *closure* of a subset  $A$  of  $M$  is the minimal span in  $M$  which contains  $A$ . Clearly, the span of  $A$ , which we always denote by  $S(A)$ , is unique. Where  $A$  is a subset of column vectors in a matrix  $M$  of column vectors,  $S(A)$  is all the columns in  $M$  which are linear combinations of  $A$ .

The terms above are used extensively in section 1.5 and section 1.6 to prove theorem 1. The terms below, through (4) and (5), are used extensively in Part II.

A subset  $A$  of  $M$  is said to *span* a subset  $K$  of  $M$  when  $K \subset S(A)$ . It follows from proposition 4, to come, that  $A$  spans  $K$  in  $M$  if and only if for each element  $e \in K$  either  $e \in A$  or there is a circuit  $C$  of  $M$  such that  $e \in C$  and  $C - e \subset A$ .

(4) The family of spanning sets of  $M$ . A *spanning subset* of  $M$  is a subset of  $M$  which spans  $M$ —in other words, a subset of  $M$  whose span is  $M$ .

(5) The family of bases of  $M$ . A *base* of  $M$  is a maximal independent set of  $M$ . A base can also be defined as a minimal spanning set of  $M$ .

The terms in (1), (2), and (5) are taken from Whitney [8]. The terms "closed set" and "span of  $A$ " are taken from Lehman [3]. There is an alternative terminology due to Tutte [7]. Since these are major sources on matroids, it is worthwhile to set down the relationship. To do so it is necessary to invoke the much used notion of "dual matroid," though it is not used here or in Part II. Papers [3], [7], and [8] show that the set-complements of the bases in a matroid  $M$  are the bases of a so-called *dual matroid*  $M^*$ .

The bases of matroid  $M$  are called by Tutte the *dendroids* of  $M$ . The elements of  $M$  are called by Tutte the *cells* of  $M$ . The independent sets of  $M$  are called by Lehman the *trees* of  $M$ .

The circuits of a matroid  $M$  are what Tutte calls the *atoms* of dual matroid  $M^*$ . The circuits of  $M^*$  are the atoms of  $M$ . Thus here is another special family of subsets of a matroid  $M$ .

(6) The family of atoms (dual circuits) in  $M$ .

The rows of a matrix  $N_0$ , under addition and subtraction, generate a group of row vectors which Tutte calls a *chain-group*, say the *chain-group*  $N$  of matrix  $N_0$ . The matroid  $M$  of matrix  $N_0$  is of course an invariant of chain-group  $N$ , and it is what Tutte calls the matroid of chain-group  $N$ . An *atom* of  $M$  of  $N$  is defined as a set of elements in  $M$  which corresponds to a minimal nonempty set of row-vector components

such that there is some member of chain-group  $N$  which has its nonzero values in precisely these components. The row-vectors orthogonal to each row of matrix  $N_0$  form another chain-group, say  $N^*$ . Its matroid is  $M^*$ , the dual of  $M$ . Atoms of  $M^*$  by definition correspond to minimal dependent sets of columns in matrix  $N_0$ . That is, they are the circuits of the matroid  $M$  of  $N_0$ .

Tutte defines a flat of matroid  $M$  to be a union of atoms of  $M$ , or the empty set. It can be shown that a flat of  $M$  is the set-complement of a span (closed set) in  $M$ , and conversely.

Where  $A$  is a subset of elements in  $M$ , Tutte denotes by  $M \cdot A$  what here is called the submatroid  $A$  of  $M$  (following Whitney). The meanings of the rank  $r(M)$  of matroid  $M$  coincide, and Tutte denotes by  $r(M \cdot A)$  what here is called the rank  $r(A)$  of set  $A$  in  $M$  (following Whitney). However, for a set  $A$ , what Tutte denotes by  $r A$  is not  $r(A) = r(M \cdot A)$  but " $r(M \times A)$ " which is used in Part II.

### 1.5. The Lemmas

In the proof of theorem 1 we will use axiom 1 and the following axiom 2' for matroids instead of axioms 1 and 2.

AXIOM 2': The union of any independent set and any element contains at most one circuit (minimal dependent set).

PROPOSITION 1: Axioms 1 and 2' are equivalent to axioms 1 and 2.

PROOF: Assuming 1 and 2, suppose independent set  $I$  together with element  $e$  contains two distinct circuits  $C_1$  and  $C_2$ . Assume  $I$  is minimal for this possibility.  $e \in C_1 \cap C_2$ . There is an element  $e_1 \in C_1 - C_2$  and an element  $e_2 \in C_2 - C_1$ . Set  $I \cup e - e_1 - e_2$  is independent since otherwise  $(I - e_1)$  is a smaller independent set than  $I$  for which  $(I - e_1) \cup e$  contains more than one circuit. Set  $I$  and set  $I \cup e - e_1 - e_2$  are maximal independent subsets of set  $I \cup e$ . This contradicts axiom 2.

Assuming 1 and 2', suppose  $I_1$  and  $I_2$  are both maximal independent subsets of a set  $A$  such that  $|I_1| < |I_2|$ . Assume  $I_1 \cup I_2$  is minimal for this possibility. There is an  $e_1$  in  $I_1 - I_2$  and  $I_2 \cup e_1$  is dependent. By 2',  $I_2 \cup e_1$  contains a unique circuit  $C$  which must contain some element  $e_2$  not in  $I_1$ . Since  $I_2$  is larger than  $I_1$  it must contain another element besides  $e_2$  not in  $I_1$  and hence  $I_2 \cup I_1 - e_2$  is dependent. Therefore, since  $I_2 \cup e_1 - e_2$  is independent, there is some  $I'_1$  such that  $e_1 \in I'_1 \subset I_1 - I_2$  and such that  $I'_1 = I_2 \cup I'_1 - e_2$  is maximal independent in  $A$ . Because  $I'_1$  contains an element not in  $I_2$ ,  $|I'_1| \geq |I_2| > |I_1|$ . However, since  $I_1 \cup I'_1$  is a proper subset of  $I_1 \cup I_2$ , this contradicts the minimality assumption for  $I_1 \cup I_2$ . The proposition is proved.

PROPOSITION 2: Axioms 1 and 2' are equivalent to the following axioms,  $1_c$  and  $2_c$ , for a matroid in terms of its circuits (where starting with circuits, independent sets are defined as sets containing no circuits).

AXIOM  $1_c$ : No circuit contains another circuit.

AXIOM  $2_c$ : If distinct circuits  $C_1$  and  $C_2$  both contain an element  $e$  then  $C_1 \cup C_2 - e$  contains a circuit.

A proof of proposition 2 is obvious.

The next very useful proposition is taken in [7] and [8] to be an axiom instead of  $2_c$ . Alfred Lehman discovered that  $1_c$  and  $2_c$  suffice.

PROPOSITION 3: If  $C_1$  and  $C_2$  are circuits of a matroid  $M$  with an element  $e \in C_1 \cap C_2$  and an element  $a \in C_1 - C_2$ , then there is a circuit  $C$  such that

$$a \in C \subset C_1 \cup C_2 - e.$$

PROOF (Lehman): Assuming  $1_c$  and  $2_c$ , suppose  $C_1, C_2, a$ , and  $e$  are such that the theorem is false and  $C_1 \cup C_2$  is minimal. There is a circuit  $C_3 \subset C_1 \cup C_2 - e$ , but  $a \notin C_3$ . There is an element  $b \in C_3 \cap (C_2 - C_1)$ . By minimality of  $C_1 \cup C_2$  for falsity of the theorem and since  $a \notin C_2 \cup C_3$ , there is a circuit  $C_4$  such that  $e \in C_4 \subset C_2 \cup C_3 - b$ . Again by the minimality and since  $b \notin C_1 \cup C_4$ , there is a circuit  $C$  such that

$$a \in C \subset C_1 \cup C_4 - e \subset C_1 \cup C_2 - e,$$

contradicting the falsity of the theorem.

PROPOSITION 4: An element  $e$  of a matroid  $M$  is in the span  $S(A)$  of a set  $A$  in  $M$  if and only if  $e$  is in  $A$  or there is a circuit  $C$  of  $M$  for which  $C - A = e$ .

PROOF: The "if" part of the theorem is asserted in the definition of span. Assuming the "only if" part false, by the definition of span there must be an  $A$  and  $e \in S(A) - A$  for which there is no  $C$  with  $C - A = e$  but for which there is a  $C$  and nonempty  $E$  with  $C - (A \cup E) = e$  where for each  $e' \in E$  there is a  $C'$  with  $C' - A = e'$ . Assume  $E$  to be minimal so that  $E \subset C$ . By prop. 3, for any  $e'$  and  $C'$  there is a  $C_1$  for which  $e \in C_1 \subset C \cup C' - e'$ . Hence,  $C_1 - (A \cup E_1) = e$  where  $E_1$  is a proper subset of  $E$ , contradicting the minimality of  $E$ .

Besides axioms 1 and 2' and the definitions of circuit and span, the only other fact on matroids used to prove theorem 1 is

PROPOSITION 5: The span of a set  $A$  in a matroid  $M$  is the (unique) maximal set  $S$  in  $M$  which contains  $A$  and which has the same rank as  $A$ .

In particular the additional fact used in proving theorem 1 is that the span of an independent set  $I$  has rank equal to the cardinality of  $I$ .

PROOF OF PROP 5: If, for  $S(A)$  the span of  $A$ ,  $r(S(A)) > r(A)$ , then by axiom 2 a base  $I$  of  $A$  is not a base of  $S(A)$ , i.e., there is an element  $e \in S(A) - I$  such that  $I \cup e$  is independent. By prop. 4,  $e$  is not in the span  $S(I)$  of  $I$  but  $A$  is in  $S(I)$ . Since the span of a set is the minimal span containing the set,  $S(A) \subset S(I)$ . Thus, by contradiction,  $r(S(A)) = r(A)$ .

Let  $e \in S'(A)$  where  $A \subset S'(A)$  and  $r(A) = r(S'(A))$ . Then, where  $I$  is a base of  $A$ , either  $e \in I$  or  $e \cup I$  is dependent. Thus  $e \in S(A)$ . Therefore,  $S(A)$  is the unique maximal set where  $A \subset S(A)$  and  $r(S(A)) = r(A)$ .

### 1.6. The Main Proof

PROOF of theorem 1 (the "if" part): Assume that for every subset  $A$  of matroid  $M$ ,  $|A| \leq k \cdot r(A)$ . Actually, it is sufficient that for every span  $S$  in  $M$ ,  $|S| \leq k \cdot r(S)$ .

The goal is to get all the elements of  $M$  into just  $k$  independent sets of  $M$ . Let  $F$  be a family of  $k$  mutually disjoint independent sets of  $M$ . Any number of these sets may be empty. These sets are to be regarded as labeled so that each may be altered in the course of the proof while still maintaining its label-identity. Suppose there is an element  $x$  of  $M$  such that  $\cup\{I_i: I_i \in F\} \subset M - x$ . We shall see how to rearrange elements among the members of  $F$  to make room for  $x$  in one of them while preserving the independence (and mutual disjointness) of them all. The process can be repeated until each element of  $M$  is in a member of  $F$ . Thus the theorem will be proved.

Implementing this proof to an algorithm for partitioning (if possible) a matroid  $M$  into  $k$  independent sets is quite straight-forward as long as an algorithm is known for the following: for any  $A \subset M$  and  $e \in M$ , find a circuit  $C$  such that  $e \in C \subset A \cup e$  or else determine that there is none. In the algorithm for partitioning  $M$ , one of course would not first verify  $|A| \leq k \cdot r(A)$  for all  $A \subset M$ , but would simply proceed on the assumption that it is true and then stop if a contradiction arises.

If every member of  $F$  contained as many as  $r(M)$  elements, then since they are disjoint and do not contain  $x$ , the union of all  $k$  of them together with  $x$ , which is a subset of  $M$ , would have cardinality greater than  $k \cdot r(M)$ . However,  $|M| \leq k \cdot r(M)$ . Hence there is an  $I_1 \in F$  for which  $|I_1| < r(M)$ . Similarly,  $x \in S_1 = S(I_1)$  implies that there is an  $I_2 \in F$  for which  $|I_2 \cap S_1| < r(S_1)$ , since if each member of  $F$  had  $r(S_1)$  elements in  $S_1$ , then their union together with  $x$  would be more than  $k \cdot r(S_1)$  elements in  $S_1$ , but  $|S_1| \leq k \cdot r(S_1)$ .

Denoting  $M$  by  $S_0$ , then likewise in general

$$x \in S_i = S(I_i \cap S_{i-1})$$

implies that there is an  $I_{i+1} \in F$  for which  $|I_{i+1} \cap S_i| < r(S_i)$ , since  $|S_i| \leq k \cdot r(S_i)$ . These  $I_i$ 's are not necessarily distinct members of  $F$ .

Where

$$S_{i+1} = S(I_{i+1} \cap S_i),$$

we have

$$r(S_{i+1}) < r(S_i).$$

Since rank is a nonnegative integer, we must eventually reach an integer  $h$  for which

$$x \notin S_h = S(I_h \cap S_{h-1})$$

and

$$x \in S_i \text{ for } i = 1, \dots, h-1.$$

By construction,  $S_1 \supset S_2 \supset \dots \supset S_h$ .

If  $I_h \cup x$  is independent then replacing  $I_h$  by  $I_h \cup x$  disposes of  $x$ . Otherwise there is a unique circuit  $C \subset I_h \cup x$ . Since  $C - x \subset S_{h-1}$  would imply  $x \in S_h = S(I_h \cap S_{h-1})$ , there is an  $x_1 \in C - x$  such that  $x_1 \notin S_{h-1}$ .

We replace  $I_h$  in  $F$  by independent set  $I_h \cup x - x_1$ . The new family is still called  $F$  and the new set carries the label-identity in  $F$  which  $I_h$  had. This and the following informal conventions are used simply to avoid introducing a lot more indices. Any other  $I_i$  which was the same member of  $F$  as  $I_h$  is now  $I_h \cup x - x_1$ . We will distinguish between the current  $I_i$  and the original  $I_i$ . The  $S_i$ 's do not change.

We have disposed of  $x$  and now we must find a place for  $x_1$  in some member of  $F$ . Since  $x_1 \notin S_{h-1}$  and  $x_1 \in S_0$ , and since the  $S_i$ 's are monotonically nested, there is some index  $i(1) \leq h-1$  for which

$$x_1 \notin S_{i(1)} \text{ and } x_1 \in S_{i(1)-1}.$$

Denote  $h$  by  $i(0)$  and denote  $x$  by  $x_0$ . Assume inductively that  $x_0 \notin S_{i(0)}$ ,  $x_0 \in S_{i(0)-1}$ ,  $x_1 \notin S_{i(1)}$ ,  $x_1 \in S_{i(1)-1}$ ,  $\dots$ ,  $x_j \notin S_{i(j)}$ ,  $x_j \in S_{i(j)-1}$ , where  $i(0) > i(1) > \dots > i(j)$ . Assume further that  $I_{i(0)}$  was replaced in  $F$  by  $I_{i(0)} \cup x_0 - x_1$ , then  $I_{i(1)}$  was replaced in  $F$  by  $I_{i(1)} \cup x_1 - x_2$ ,  $\dots$ , and then  $I_{i(j-1)}$  was replaced in  $F$  by  $I_{i(j-1)} \cup x_{j-1} - x_j$ ; where  $x_1 \in C_0 \subset I_{i(0)} \cup x_0$ ,  $x_2 \in C_1 \subset I_{i(1)} \cup x_1$ ,  $\dots$ , and  $x_j \in C_{j-1} \subset I_{i(j-1)} \cup x_{j-1}$ .

Suppose there is a circuit  $C_j \subset I_{i(j)} \cup x_j$ . Set  $I_{i(j)}$  might have the same label-identity in  $F$  as  $I_{i(q)}$  for several values of  $q < j$ , and so the contents of  $I_{i(j)}$  may have changed several times since the original  $I_{i(j)}$  which gave rise to  $S_{i(j)} = S(I_{i(j)} \cap S_{i(j)-1})$ . In particular,  $x_q$  for some  $q < j$  may have been adjoined to  $I_{i(j)}$ . However, by the induction hypothesis any such  $x_q$  is contained in  $S_{i(q)-1}$  and thus in  $S_{i(j)}$ .

Therefore all elements of  $C_j - x_j$  which are not in the original  $I_{i(j)}$  are in  $S_{i(j)}$ . By definition of  $S_{i(j)}$ , all elements of the original  $I_{i(j)}$  which are in  $S_{i(j)-1}$  are also in  $S_{i(j)}$ . Thus if all elements of  $C_j - x_j$  are in  $S_{i(j)-1}$  then they are all in  $S_{i(j)}$ , but since  $S_{i(j)}$  is a span then  $x_j$  also would be in  $S_{i(j)}$ , contradicting the inductive hypothesis. Hence, there exists some element  $x_{j+1}$  of  $C_j$  such that  $x_{j+1} \notin S_{i(j)-1}$ . Since  $x_{j+1} \in S_0$ , there is some  $i(j+1) < i(j)$  such that  $x_{j+1} \notin S_{i(j+1)}$  and  $x_{j+1} \in S_{i(j+1)-1}$ .

Therefore when there exists a  $C_j$ , we repeat the inductive step by replacing  $I_{i(j)}$  by  $I_{i(j)} \cup x_j - x_{j+1}$ .

Since  $i(0) > i(1) > \dots$ , eventually we must reach an  $i(j)$  for which there is no  $C_j \subset I_{i(j)} \cup x_j$ . Then we can replace  $I_{i(j)}$  in  $F$  by independent set  $I_{i(j)} \cup x_j$  without having to displace another element  $x_{j+1}$ . End of proof.

### 1.7. Corollary

For the special case where  $M$  is the matroid of a graph  $G$ , theorem 1 can be simplified somewhat:

COROLLARY (Nash-Williams [5]): *The edges of a graph  $G$  can be colored with as few as  $k$  colors so that no circuit of  $G$  is all one color, if and only if there is no subset  $U$  of nodes in  $G$  such that, where  $E_U$  is the set of edges in  $G$  which have both ends in  $U$ ,*

$$|E_U| > k(|U| - 1).$$

Symbols  $|U|$  and  $|E_U|$  denote, respectively, the cardinalities of  $U$  and  $E_U$ .

Not every subset  $A$  of elements in the matroid  $M(G)$  of  $G$ , nor even every closed set  $A$  of elements in  $M(G)$ , corresponds to a set of edges of type  $E_U$ . However, the relation  $|A| \leq k \cdot r(A)$  for every set  $A$  corresponding to a set  $E_U$  of edges which form a connected subgraph of  $G$  implies the relation for every subset  $A$  of elements in  $M(G)$ .

The corollary follows (we omit the proof) from theorem 1 by using the following characterization of the rank function of a graph due to Whitney:

The rank  $r(E)$  of any subset  $E$  of edges in  $G$ , i.e., the rank of the matroid subset corresponding to  $E$ , equals the number of nodes minus the number of connected components in the subgraph,  $G \cdot E$ , consisting of the edges  $E$  and the nodes they meet, or equivalently in the subgraph,  $G : E$ , consisting of the edges  $E$  and all the nodes of  $G$ . The notation  $G \cdot E$  and  $G : E$  is due to Tutte, chapter III of [7].

## References

- [1] G. Birkhoff, Abstract linear dependence and lattices, *Amer. J. Math.* 57, 800-804 (1935).
- [2] H. Crapo, Single-element Extensions of Matroids, *J. Res. NBS* 69B (Math. and Math. Phys.) No. 1.
- [3] A. Lehman, A solution of the Shannon switching game, Univ. of Wisc. Math. Research Center Report #308, 1962, to appear in the *SIAM Journal*.
- [4] C. St. J. A. Nash-Williams, Edge-disjoint spanning trees of finite graphs, *J. London Math. Soc.* 36, 445-450 (1961).
- [5] C. St. J. A. Nash-Williams, Decomposition of finite graphs into forests, *J. London Math. Soc.* 39, 12 (1964).
- [6] W. T. Tutte, On the problem of decomposing a graph into  $n$  connected factors, *J. London Math. Soc.* 36, 221-230 (1961).
- [7] W. T. Tutte, Lectures on matroids, *J. Res. NBS* 69B (Math. and Math. Phys.) No. 1.
- [8] H. Whitney, On the abstract properties of linear dependence, *Amer. J. Math.* 57, 509-533 (1935).

(Paper 69B1-134)

# Lehman's Switching Game and a Theorem of Tutte and Nash-Williams<sup>1</sup>

Jack Edmonds

(December 1, 1964)

The results cited in the title are unified by the following theorem: For any matroid  $M$  and any subsets  $N$  and  $K$  of elements in  $M$ , there exist as many as  $k$  disjoint subsets of  $N$  which span  $K$  and which span each other if and only if there is no contraction matroid  $M \times A$  where  $N \cap A$  partitions into as few as  $k$  sets such that each is independent in  $M \times A$  and such that at least one of them does not span  $K \cap A$  in  $M \times A$ .

## 2.1. The Problem

A. Lehman [3]<sup>2</sup> posed the following game to be played between two players on any given matroid  $M$  with a distinguished element  $e$ . The players are called the cut player and the short player. They take turns and (to be explicit) the cut player goes first. Each player in his turn tags an element of  $M$ , other than  $e$ , not already tagged. The short player wins if he tags a set of elements which span  $e$ . The cut player wins otherwise—that is, the cut player wins if the elements, other than  $e$ , which he has not tagged do not span  $e$ .

The game, determined by  $M$  and  $e$ , is called a *short game* if the short player can win against any strategy of the cut player. We will call the game *nonshort* if the cut player can win against any strategy of the short player. Clearly a game is one or the other. For any  $M$  and  $e$ , Lehman characterizes short games and describes a winning strategy for the short player.

Recall from section 1.4 that a set  $T$  in a matroid  $M$  is said to span a set  $A$  in  $M$  if for every  $e \in A$ , either  $e \in T$  or there is a circuit  $C$  of  $M$  such that  $C - e \subset T$ . Recall that a base  $B$  of  $M$  is a set which spans  $M$  (e.g., a spanning set of  $M$ ) and which also is independent.

Where the game is played on a graph  $G$ , it is not necessary to have an edge corresponding to  $e$  but sufficient to have two distinguished "terminal" nodes,  $v_1$  and  $v_2$ , which would be the ends of  $e$ . Here, the goal of the short player is to tag a set of edges which contains a path of edges joining  $v_1$  to  $v_2$ . The goal of the cut player is to tag a set of edges which separates  $v_1$  from  $v_2$ .

A theorem due independently to Tutte [6] and Nash-Williams [4] characterizes for any graph  $G$  the maximum number of edge-wise disjoint subgraphs, each connected and containing all nodes of  $G$ , into which the edges of  $G$  can be partitioned. For a connected graph  $G$ , the edges of a connected subgraph which contains every node of  $G$  correspond to the elements of a spanning set of the matroid of  $G$ , and conversely.

The purpose of the present note is to unify these two theories. Theorem 2 states the straightforward generalization to matroids of the Tutte and Nash-Williams theorem. Theorem 3 is Lehman's main theorem characterizing short games. Theorem 4 is an analogous theorem characterizing nonshort games. (Lehman characterizes nonshort games indirectly by using "dual matroids" which we avoid.) Theorem 5, for the case where  $K = N = M$ , yields theorem 2. For the case where  $k = 2$ , it yields the "only if" parts of theorems 3 and 4. The "if" parts of theorems 3 and 4 are proved by describing the winning strategies when the respective conditions hold (in one case this follows Lehman, [3]).

Theorem 1 in section 1.3 and theorem 2 are in a sense dual to each other but not in the usual matroid sense. Each can be proved from the other. We use theorem 1 to prove theorem 5.

Theorem 5 appears interesting in itself. We call it the "cospanning-set theorem" after a main idea of Lehman's theory. For a graph  $G$  with a prescribed subset of nodes called terminals, it gives a "good" characterization for the nonexistence of  $k$  edge-wise disjoint connected subgraphs (e.g., trees), all with precisely the same set of nodes which includes the terminals.

If the matroid  $M$  of the cospanning-set theorem is a finite set of vectors in a space  $L$ , then for given subsets  $N$  and  $K$  of  $M$ , the theorem provides a "good" characterization for the nonexistence of as many as  $k$  disjoint subsets  $N_i$  of  $N$  and a subspace  $L'$  of  $L$  such that each  $N_i$  exactly spans  $L'$  and such that  $L'$  contains  $K$ .

## 2.2. Contractions

We use the following important concept on matroids due to Tutte (ch. II of [7]). For any set  $A$  of elements in a matroid  $M$ , define the circuits of  $M \times A$  to be the minimal nonempty intersections of  $A$  with circuits of  $M$ .

PROPOSITION 6: The set of elements  $A$  and the circuits of  $M \times A$  are a matroid (denoted by  $M \times A$ ), called the contraction of  $M$  to  $A$ .

PROOF: Axioms 1<sub>c</sub> and 2<sub>c</sub> for  $M \times A$  follow immediately from prop. 3 for  $M$ .

<sup>1</sup> This paper is a sequel to the preceding one, "Minimum Partition of a Matroid Into Independent Subsets." The numbering system there, including references, is continued here. This work was supported by the Army Research Office (Durham) and the Defense Communications Agency.

<sup>2</sup> Figures in brackets indicate the literature references on page 72.

**COROLLARY:** Where  $A$  and  $\bar{A}$  are complementary subsets of matroid  $M$ ,  $\bar{A}$  is closed (a span) in  $M$  if and only if matroid  $M \times A$  contains no "loops," that is elements of rank zero.

**PROPOSITION 7:** Where  $K$  and  $A$  are subsets of matroid  $M$ , subset  $T'$  of  $A$  spans  $K \cap A$  in  $M \times A$  if and only if there is a subset  $T$  of  $M$  such that  $T' = T \cap A$  and such that  $T$  spans  $K$  in  $M$ .

**COROLLARY:** The spanning sets of matroid  $M \times A$  are precisely the intersections of  $A$  with spanning sets of  $M$ .

**PROOF OF PROP. 7:** Suppose  $T' = T \cap A$  where  $T$  spans  $K$  in  $M$ . Since  $T$  spans  $K$ , for any element  $e$  in  $K \cap A$ , either  $e \in T$  or there is a circuit  $C$  in  $M$  such that  $C - e \subset T$ . If  $e \in T$  then  $e \in T'$  and hence  $T'$  spans  $e$  in  $M \times A$ . If there is a  $C$  then, by definition of  $M \times A$ , there is a circuit  $C'$  of  $M \times A$  such that  $e \in C' \subset C$ . It follows that  $C' - e \subset T'$  and hence  $T'$  spans  $e$  in  $M \times A$ . Thus, the "if" part is proved.

Suppose subset  $T'$  of  $A$  spans  $K \cap A$  in  $M \times A$ . Let  $T = T' \cup \bar{A}$  where  $\bar{A}$  is the complement of  $A$  in  $M$ . Then  $T' = T \cap A$ . Let  $e$  be any element of  $K$ . If  $e \in T$ , then  $T$  spans  $e$  in  $M$ . Otherwise,  $e \in K \cap A$ , and  $e \in T'$ . Since  $T'$  spans  $e$  in  $M \times A$ , there is a circuit  $C'$  of  $M \times A$  such that  $e \in C'$  and  $C' - e \subset T'$ . By definition of  $M \times A$ , there is a circuit  $C$  of  $M$  such that  $C' = C \cap A$ . Therefore,  $T$  spans  $e$ , since  $e \in C$  and  $C - e \subset T$ . Thus, the "only if" part is proved.

Tutte uses  $M \cdot A$  to denote what we mean by the submatroid  $A$  of  $M$ ; he does not follow Whitney's informality of letting  $A$  mean both a matroid and its set of elements. We will use Tutte's notation and also, where convenient, we will depart from it again by referring to  $M \times A$  simply as the contraction matroid,  $A$ , of  $M$  just as we refer to  $M \cdot A$  as the submatroid,  $A$ , of  $M$ . Also,  $A$  denotes the elements of either.

Where  $M(G)$  is the matroid of graph  $G$ , the matroid of a subgraph  $H$  of a graph  $G$  is the submatroid of  $M(G)$  which contains the elements corresponding to the edges of  $H$  and conversely. The matroid of a "contraction graph"  $H$  of  $G$  is the contraction of  $M(G)$  which contains the elements corresponding to the edges of  $H$ , and conversely.

The most instructive way to describe the meaning of contraction graph is visually. The contraction graph  $H$  of  $G$  whose edges are the set  $H$  of edges in  $G$  is the graph obtained from  $G$  by contracting to a point each edge of  $G$  not in  $H$ .

It should be pointed out that in order for there to be a contraction  $H$  of  $G$  for every subset  $H$  of edges in  $G$ , we must extend our meaning of graph (in sec. 1.1) to graphs which include edges which "meet the same node at both ends." These "loop" edges are circuits by themselves; they correspond to matroid elements which are not contained in any independent set of the matroid. This sort of matroid element corresponds in a matrix to a column of all zeros. In a matrix  $N(G)$ , a loop of graph  $G$  can be represented by a column of  $N(G)$  which contains a 2 in the row corresponding to the node met and which contains zeros elsewhere. Relative to the matroid structure, the column is all zeros, mod. 2.

We have pointed out how any contraction of the matroid of a graph can be represented as the matroid of a graph. It is also possible to represent any contraction of the matroid of a matrix as the matroid of a matrix.

By deleting (or cutting) set of elements  $A$  in matroid  $M$ , we mean replacing matroid  $M$  by its submatroid on the set  $M - A$ . By contracting (or shorting) set of elements  $A$  in matroid  $M$ , we mean replacing matroid  $M$  by its contraction to the set  $M - A$ . Clearly, from the definition of submatroid, we can get a submatroid  $M'$  of  $M$  by deleting the elements of  $M$  not in  $M'$  one after another in any order. Clearly, from the corollary to prop. 7, we can get a contraction matroid  $M'$  of  $M$  by contracting the elements of  $M$  not in  $M'$  one after another in any order. It can be proved that for any elements  $a$  and  $b$  in a matroid  $M$ , deleting  $a$  and then contracting  $b$  is the same as contracting  $b$  and then deleting  $a$ . The proof is omitted. These results can be summarized by the following:

**PROPOSITION 8:** The operations of deleting certain elements together with the operations of contracting certain other elements in a matroid are associative and commutative.

The above proposition is equivalent to Tutte's identities 3.33 in [7]. Tutte defines a *minor* of a matroid  $M$  to be any matroid obtained from  $M$  by deleting certain elements and contracting certain other elements in  $M$ .

The following theorem is presented by Tutte (theorem 3.53 of [7]) in terms of "dendroids."

**PROPOSITION 9:** If  $A$  and  $\bar{A}$  are complementary sets of elements in matroid  $M$ , then the elements in a base of  $M \times A$  together with the elements in a base of  $M \cdot \bar{A}$  are the elements in a base of  $M$ .

Proof omitted.

**COROLLARY:**  $r(M \cdot A) + r(M \times \bar{A}) = r(M)$ .

We have been calling  $r(M \cdot A)$  the rank  $r(A)$  of set  $A$  in matroid  $M$ . We denote  $r(M \times A)$  as function  $r(A)$  of sets  $A$  in matroid  $M$ .

The following theorem, which for the case of connected graphs is the one due to Tutte and Nash-Williams, completely parallels theorem 1. The "if" part of theorem 2 follows immediately from the "if" part of theorem 5 (where  $M = N = K$ ).

**THEOREM 2:** The elements of a matroid  $M$  can be partitioned into as many as  $k$  sets, each a spanning set of  $M$ , if and only if there is no subset  $A$  of elements of  $M$  for which

$$|A| < k \cdot r(A).$$

Any contraction graph of a connected graph is connected. Using the last paragraph of 1.7, observe that where  $M$  is the matroid of a connected graph  $G$ ,  $r(A)$  is the number of nodes minus one of a contraction graph of  $G$ , and  $|A|$  is the number of edges in that contraction graph.

Notice that, since  $r(A) = r(M) - r(\bar{A})$ , theorem 2 is easily stated without the notion of contraction.

To prove the "only if" part of theorem 2, assume that  $M$  partitions into  $k$  sets, each spanning  $M$ . By taking a subset of each of them, we get disjoint bases

$B_i (i = 1, \dots, k)$ . Let  $A$  be any subset of  $M$  and let  $\bar{A}$  be its complement. Since  $B_i$  is independent,  $r(\bar{A}) \geq |\bar{A} \cap B_i|$ . Since  $B_i$  is a base,  $|A \cap B_i| + |\bar{A} \cap B_i| = |B_i| = r(M)$ . Combining the two gives  $|A \cap B_i| \geq r(M) - r(\bar{A}) = r(A)$ . Therefore  $|A| \geq \sum_i |A \cap B_i| \geq k \cdot r(A)$ .

### 2.3. Short Games

It turns out to be just as easy to analyze games where, for the graph case, any subset of nodes of  $G$  are distinguished as terminals and the goal of the short player is to tag a set of edges in  $G$  which contains the edges of a connected subgraph containing all the terminals. To interpret this game in matroid terms, adjoin to  $G$  a set of new edges which form a connected graph  $K$  containing precisely the terminals as nodes. Then relative to the matroid of graph  $G \cup K$ , the goal of the short player is to tag a set of elements corresponding to edges in  $G$  which spans the set of elements corresponding to edges in  $K$ .

For any matroid  $M$  and nonempty subsets  $N$  and  $K$ , consider the game  $L(M, N, K)$  where, as before, the cut player and short player take turns tagging different elements of  $N$ , the cut player going first. The short player wins if he tags a set of elements which span  $K$ . Otherwise, the cut player wins. Call  $L(M, N, K)$  a short game if the short player can win against any strategy of the cut player.

Lehman's main theorem (explicitly for the case where  $K$  is a single element) is

**THEOREM 3:**  $L(M, N, K)$  is a short game if and only if  $N$  contains two disjoint sets,  $A_0$  and  $B_0$ , of elements which span each other and which span  $K$ .

Notice that in the two-terminal graph case, the short player wants to get a path joining the terminals. The structure characterizing when he can is two edgewise disjoint trees each containing the terminals and each containing precisely the same nodes as the other.

Lehman calls two (or more) sets which span each other *cospanning*. Let us verify that two disjoint cospanning sets  $A_0$  and  $B_0$  in  $N$  which span  $K$  provide a winning strategy in the game  $L(M, N, K)$  for the short player. All that we need consider is the span  $M_0 = S(A_0) = S(B_0)$  in  $M$ . Clearly, we can take  $A_0$  and  $B_0$  to be bases of submatroid  $M_0$ ; assume that they are. If the cut player tags an element not in  $A_0 \cup B_0$ , we can pretend that at the same time he also tags some element of  $A_0 \cup B_0$ . Clearly, the short player would not be taking an illegal advantage by pretending this. Therefore, suppose the cut player in his first turn tags element  $a_0$  in  $A_0$ .

By axiom 2 (in 1.1) there is an element  $b_0$  of  $B_0$  such that  $(A_0 - a_0) \cup b_0$  is a base of  $M_0$ . The short player should tag an element  $b_0$ . It follows from prop. 7 that disjoint sets  $A_1 = A_0 - a_0$  and  $B_1 = B_0 - b_0$  are spanning sets of the contraction matroid  $M_1 = M_0 - b_0$  of  $M_0$ .

Since it is the cut player's turn again, the situation of  $A$  and  $B$  relative to  $M_1$  is as it was for  $A_0$  and  $B_0$  relative to  $M_0$  except that  $M_1$  is smaller. Assuming there is a strategy for the succeeding turns whereby the short player can tag a set of elements which contains a base  $T$  of reduced matroid  $M_1$ , then by prop. 9

the set  $T \cup b_0$  of elements, which the short player will have tagged, is a base of matroid  $M_0$  and hence spans set  $K$  in matroid  $M$ .

When  $B_0$  contains only one element  $b_0$ , then  $b_0$  itself spans  $M_0$  and  $K$ . Hence, by induction on the number of elements, we have a winning strategy for the short player. This proves the "if" part of theorem 3. The harder "only if" part will follow from theorem 4 and theorem 5.

### 2.4. Nonshort Games

The notion of contraction can always be used in place of the more familiar notion of "matroid duality," and conversely, because of a theorem (3.27 of [7]) relating the contraction matroids of an  $M$  to the submatroids of the "dual to  $M$ ." Sometimes one notion is convenient, sometimes the other. We do not use duality here. Lehman in treating the same topic uses mainly duality.

Lehman's interpretation of his dual results characterizing when the cut player can win for the case of graphs does not directly provide a "good" characterization in the sense of the absolute supervisor. Clearly his characterization of a short game is good in the case of graphs. However, he does not give the following analogous characterization for nonshort games. (Compare Lehman's theorem (26) and its graph interpretation with our theorem 4 and its contraction graph interpretation. See also the comment on his theorem (26) which follows his theorem (29).)

**THEOREM 4:**  $L(M, N, K)$  is a non-short game if and only if there is a contraction matroid  $M'$  of matroid  $M$  where set  $N' = N \cap M'$  can be partitioned into two sets  $I_1$  and  $I_2$  such that  $I_1$  and  $I_2$  are both independent in  $M'$  and such that  $I_2$  does not span the set  $K' = K \cap M'$  in  $M'$ .

Let us verify that an  $M'$ ,  $I_1$ , and  $I_2$  provide a winning strategy for the cut player in game  $L(M, N, K)$ . If  $I_1$  does not span  $K'$  in  $M'$  then the cut player can tag anything on his first turn. Otherwise, he should tag an element  $e_1$  in  $I_1$  such that  $I_1 - e_1$  does not span  $K'$  in  $M'$ .

Since  $I_2$  does not span  $K'$ , there is an element  $e_2 \in K'$  such that  $r(e_2) \neq 0$ . If  $e_2 \in I_1$ , then  $e_1$  is an element  $e_1$ . Otherwise, by axiom 2' there is a unique circuit  $C$  in  $I_1 \cup e_2$  and so any element of  $C - e_2$  is an element  $e_1$ . ( $C - e_2$  is not empty since  $r(e_2) \neq 0$ .) Now neither the untagged elements  $I'_1 = I_1 - e_1$  of  $I_1$  nor the untagged elements  $I'_2 = I_2$  of  $I_2$  span  $K'$  in contraction matroid  $M'$ .

Even if the short player tags an element not in  $M'$ , clearly the cut player is not taking an illegal advantage by pretending the short player also tags an untagged element in  $M'$  if there are any. Therefore, assume the short player does tag one, say  $e_2$  in  $I'_2$ . Consider the contraction matroid  $M'' = M' - e_2$  of  $M'$ . (By prop. 8, the contraction of matroid  $M'$  to set  $M''$  is the same as the contraction of matroid  $M$  to set  $M''$ .) By the circuit definition of contraction matroid, set  $I'_1 = I'_1 - e_2$  will be independent in  $M''$  and will not span  $K' = K' \cap M''$  in  $M''$ .

Again by the definition of contraction matroid, if  $e_2$  is not in the span of  $I'_1$  in  $M'$  then  $I'_1$  is independent in  $M''$ . In this case, the cut player should tag some ele-

ment  $e_i$  such that  $I_i' = I_i - e_i$  does not span  $K$  in  $M'$ . By the definition of contraction matroid and prop. 3, if  $e_2$  is in the span of  $I_1$  in matroid  $M'$  then set  $I_1'$  contains just one circuit of contraction  $M'$  and does not span  $K$  in  $M'$ . In this case, the cut player should tag some element  $e_i$  in the one circuit of  $I_1$  in  $M'$ , so that  $I_1' = I_1 - e_i$  is independent in  $M'$ .

Thus in either case, after the cut player takes his second turn, the untagged elements of  $M'$  partition into sets  $I_1'$  and  $I_2'$  where, in contraction  $M'$ , both are independent and neither spans  $K$ . There are no elements tagged by the short player in  $M'$ . The situation is identical to the one in  $M'$  right after the cut player took his first turn except that  $M'$  has fewer untagged elements.

Hence, by induction on the number of untagged elements in the contraction matroid, if the cut player tags as described, he eventually reaches a contraction matroid  $M^{(k)}$  in which all the elements are tagged by him, and yet for which there is an  $e \in K^{(k)} = M^{(k)} \cap K$  such that  $r(e) \neq 0$  in  $M^{(k)}$  (since  $K^{(k)}$  is not spanned by the empty  $I_1^{(k)}$  or the empty  $I_2^{(k)}$ ). The cut player will then have won the game, because for the short player to win he must tag a set, say  $T$ , which spans  $K$  in matroid  $M$ . By prop. 7, for any such  $T$  and any set  $M^{(k)}$  in  $M$ ,  $T \cap M^{(k)}$  must span  $K^{(k)} = K \cap M^{(k)}$  in matroid  $M^{(k)}$ , which is impossible. This proves the "if" part of theorem 4.

### 2.5. Cospinning-Sets Theorem

We still have to prove the "only if" parts of theorems 3 and 4. They follow immediately from theorem 5 (for the case  $k=2$ ). We proved the part of theorem 3 which says " $P \Rightarrow (L \text{ is a short game})$ ". We proved the part of theorem 4 which says " $Q \Rightarrow (L \text{ is not a short game})$ ". Theorem 5 says " $P \Leftrightarrow \text{not } Q$ ". Logic yields that " $(L \text{ is a short game}) \Rightarrow P$ " and " $(L \text{ is not a short game}) \Rightarrow Q$ ".

**THEOREM 5:** For any matroid  $M$  and any subsets  $N$  and  $K$  of elements in  $M$ , there exist as many as  $k$  disjoint subsets of  $N$  which span each other and which span  $K$ , if and only if there is no contraction matroid  $M'$  of  $M$  where  $N \cap M'$  partitions into as few as  $k$  sets such that each is independent in  $M'$  and such that at least one of them does not span  $K \cap M'$ .

**PROOF:** The "only if" part of theorem 5 follows from the "if" part of prop. 7. Suppose in matroid  $M$  there exist  $k$  disjoint subsets  $T_i$  of  $N \cap M$ , which span each other and which span  $K \cap M$ . Let  $M'$  be any contraction of  $M$ . Where a set  $T_i$  is the  $T$  of prop. 7; where set  $M'$  is the  $A$  and matroid  $M'$  is the  $M \times A$  of prop. 7; and where  $S_i$ , the span (closure) in  $M'$  of each  $T_i$ , is the  $K$  of prop. 7; prop. 7 says that  $T_i' = T_i \cap M'$  spans  $S_i' = S_i \cap M'$  in matroid  $M'$ . Since each  $T_i$  spans  $S_i$ , each  $T_i'$  contains at least  $r(S_i)$  elements where  $r(S_i)$  is the rank of set  $S_i$  in matroid  $M'$ . Since all the sets  $T_i'$  are mutually disjoint,  $N \cap S'$  contains at least  $k \cdot r(S')$  elements.

On the other hand, suppose  $N \cap M'$  partitions into as few as  $k$  independent sets  $I_i$  of  $M'$  where one of

them  $I_1$  does not span  $K \cap M'$  and hence does not span  $S'$ . Since each  $I_i' = I_i \cap S'$  is independent, each  $I_i'$  contains at most  $r(S')$  elements. Since  $I_1'$  does not span  $S'$ , it contains fewer than  $r(S')$  elements. Therefore  $N \cap S' = \cup I_i'$  contains fewer than  $k \cdot r(S')$  elements. Thus, the "only if" part of theorem 5 is proved.

The "if" part of theorem 5 follows from propositions 8 and 9 and theorem 1. Let  $M$  be any matroid, let  $N$  and  $K$  be any subsets of  $M$ , and let  $k$  be any positive integer. Suppose  $A_0$  is a maximal subset of  $N$  such that  $|A_0| = k \cdot r(A_0)$  and  $|A| \leq k \cdot r(A)$  for all  $A \subset A_0$ . Set  $A_0$  may be empty. By theorem 1,  $A_0$  partitions into  $k$  independent sets,  $I_i$ . Since  $|A_0| = k \cdot r(A_0)$ , each  $I_i$  must be a base of submatroid  $A_0$  and of course also a base of  $St(A_0)$ , the span of  $A_0$  in  $M$ .

Let  $M'$  be the contraction matroid of  $M$  obtained by contracting  $St(A_0)$  in  $M$ . Suppose  $A_1$  is a subset of  $N' = N \cap M'$  such that  $|A_1| = k \cdot r(A_1)$  and  $|A| \leq k \cdot r(A)$  in matroid  $M'$  for all  $A \subset A_1$ . Then like  $A_0$  in  $M$ ,  $A_1$  partitions into  $k$  bases  $I_i'$  of submatroid  $A_1$  of  $M'$ . By prop. 8, submatroid  $A_1$  of  $M'$  is the contraction to  $A_1$  of the submatroid  $A_1 \cup St(A_0)$  of  $M$ . Call it minor  $A_1$ . By prop. 9, a base of minor  $A_1$  together with a base of submatroid  $St(A_0)$  of  $M$  is a base of submatroid  $A_1 \cup St(A_0)$  of  $M$ .

In particular, by pairing the sets  $I_i'$  one-to-one with the sets  $I_i$ , we get  $k$  disjoint bases  $I_i' = I_i \cup I_i$  of submatroid  $A_1 \cup St(A_0)$  of  $M$ . Since  $\cup I_i' = A_0 \cup A_1 \subset N$  and since the sets  $I_i'$  span each other in  $M$ ,  $|A_0 \cup A_1| = k \cdot r(A_0 \cup A_1)$  and, by the "only if" part of theorem 1,  $|A| \leq k \cdot r(A)$  in  $M$  for all  $A \subset A_0 \cup A_1$ . However,  $A_0$  was taken to be maximal for this property, and hence  $A_1$  is empty. Thus, matroid  $M'$  contains no nonempty  $A_1$  as defined.

Since  $St(A_0)$  is closed in  $M$ , the matroid  $M'$  obtained by contracting  $St(A_0)$ , contains no element of rank zero (corollary to prop. 6). Suppose  $N' = N \cap M'$  contains a nonempty set  $A_2$  such that  $|A_2| \geq k \cdot r(A_2)$  in  $M'$ . Take  $A_2$  to be minimal. By the nonexistence in  $M'$  of a nonempty  $A_1$  as described above, we have that  $|A_2| > k \cdot r(A_2)$ . Since there are no elements of rank zero,  $A_2$  contains at least two elements. Deleting an element from  $A_2$  to get a nonempty  $A_3$ , we have  $|A_3| \geq k \cdot r(A_3) \geq k \cdot r(A_2)$  in  $M'$ , which contradicts the minimality of  $A_2$ . Therefore, for all nonempty subsets  $A$  of  $N'$ ,  $|A| \leq k \cdot r(A)$  in  $M'$ .

Suppose some element  $g \in K$  is contained in matroid  $M'$ . Since  $g$  does not have zero rank, there exists a matroid  $M_A$  which contains the elements of  $M'$  plus a new auxiliary element  $h$ , such that  $h$  and  $g$  form a circuit in  $M_A$  and such that submatroid  $M_A - h$  of  $M_A$  is the matroid  $M'$ . It is easy to verify that  $M_A$  is such a matroid where the circuits of  $M_A$  are (1) the set consisting of  $g$  and  $h$ , (2) the circuits of  $M'$ , and (3) sets  $(C - g) \cup h$  where  $C$  is a circuit of  $M'$  which contains  $g$ . Let  $N_A = N' \cup h$ . It follows from the relation  $|A| \leq k \cdot r(A)$  in matroid  $M'$  for all nonempty  $A \subset N'$ , that  $|A_A| \leq k \cdot r(A_A)$  in  $M_A$  for all  $A_A$  in  $N_A$ .

Hence by theorem 1,  $N_A$  can be partitioned into  $k$  independent sets  $I_i^A$  of  $M_A$ , including the set, say  $I_1^A$ , which contains  $h$ . In matroid  $M'$  the set  $I_1^A - h$  is independent and does not span  $g$ . All of the other sets  $I_i^A$  are independent in  $M'$ . These sets  $I_1^A$  and  $I_i^A - h$  are a partition of  $N'$ .

Thus, if there is no such partition of  $N' = N \cap M'$  for contraction  $M'$  of  $M$  then no element of  $K$  is in  $M'$ . Thus  $K \subset S(A_0)$ . In this case, the  $k$  bases  $I_i$  of submatroid  $S(A_0)$  of  $M$  span each other and span  $K$  in  $M$ . This completes the proof of theorem 5.

(Paper 69B1-135)

# Transversals and Matroid Partition\*

Jack Edmonds and D. R. Fulkerson

(June 9, 1965)

In section 1, *transversal matroids* are associated with "systems of distinct representatives" (i.e., transversals) and, more generally, *matching matroids* are associated with matchings in graphs. The transversal matroids and a theorem of P. J. Higgins on disjoint transversals of a family of sets, along with the well-known graphic matroids and some theorems on decomposition of graphs into forests, motivate some theorems on partitions of general matroids into independent sets. In section 2, the relationship between transversal result and matroid result is illustrated for a special case of later theorems. In section 3, theorems on transversals are proved using network flows. In sections 4 and 5, theorems on matroids are presented which imply various results on decomposition into transversals or into forests. In section 6, the matching matroids are shown to be simply the transversal matroids. For the most part, sections 2, 3, 4–5, and 6 can be read separately.

## 1. Transversal Matroids

A *matroid*  $M = (E, F)$  is a finite set  $E$  of elements and a family  $F$  of subsets of  $E$ , called *independent sets*, such that (1) every subset of an independent set is independent; and (2) for every set  $A \subset E$ , all maximal independent subsets of  $A$  have the same cardinality, called the *rank*  $r(A)$  of  $A$ .

Sometimes no explicit distinction is made between a matroid and its set of elements, in the same way that no explicit distinction is made between groups, spaces, or graphs and their sets of members. For example, one normally uses the same symbol to denote a space and the set of points in a space. On the other hand, it is often desirable to consider various matroids that have the same set of elements.

The primary example of a matroid is obtained by letting  $E$  be the set of columns in a matrix over some field and  $F$  the family of linearly independent subsets of columns. In particular,  $E$  may be the set of edges in a graph and  $F$  the family of edge-sets that comprise "forests" in the graph. A matroid that is abstractly isomorphic to one of the latter kind is called *graphic*.

Our motivation here will be another source of matroids, which is an extensive theory in its own right. It is well known in various contexts, including systems of distinct representatives, (0, 1)-matrices, network flows, matchings in graphs, marriages, and so forth (see [3]).<sup>1</sup> Here we will refer to it very broadly as transversal theory.

Let  $Q = \{q_i; i = 1, \dots, m\}$  be a family of (not necessarily distinct) subsets of a set  $E = \{e_j; j = 1, \dots, n\}$ . The set  $T = \{e_{j(1)}, \dots, e_{j(t)}\}$ ,  $0 \leq t \leq n$ , is called a *partial transversal* (of size  $t$ ) of  $Q$  if  $T$  consists of distinct elements in  $E$  and if there are distinct integers  $i(1), \dots, i(t)$  such that  $e_{j(k)} \in q_{i(k)}$  for  $k = 1, \dots, t$ . The set  $T$  is called a *transversal* or a *system of distinct representatives* of  $Q$  if  $t = m$ .

**THEOREM.** Let  $Q$  be any finite family of (not necessarily distinct) subsets of a finite set  $E$ . (a) If  $F$  is the family of partial transversals of  $Q$ , then  $M_Q = (E, F)$  is a matroid. (b) If  $F$  is the collection of subfamilies of  $Q$  that have transversals, then  $M_Q = (Q, F)$  is a matroid.

The statements (a) and (b) are equivalent and refer to the same abstract class of matroids because the roles of  $Q$  and  $E$  are actually symmetric. The situation is easily visualized in the form of the "incidence graph" of  $(E, Q)$ : a "bipartite" graph,  $G = G(E, Q)$ , where the nodes in one part are members of  $Q$  and the nodes in the other part are the members of  $E$ . The edges of  $G$ , which all go from one part to the other, are the incidences between  $Q$  and  $E$ .

A *transversal matroid* is one that is abstractly isomorphic to an  $M_Q$  (or an  $M_Q$ ). Matroid theory and transversal theory enhance each other via transversal matroids, as do matroid theory and graph theory via graphic matroids.

Let  $E$  be any fixed subset of nodes in any given graph  $G$ . We assume throughout this paper that each edge of a graph meets two distinct nodes. Let subset  $T \subset E$  be a member of  $F$  when  $T$  meets (is contained in the set of endpoints of) some matching in  $G$ . (A *matching* in a graph is a set of its edges such that no two members of the set meet the same node.) We shall show that  $M_{G, E} = (E, F)$  is a matroid by verifying axiom (2). In general, where  $G$  is not necessarily bipartite and where  $E$  is any subset of nodes, we call  $M_{G, E}$  a *matching matroid*. For any  $A \subset E$ , let  $T_1$  and  $T_2$  be maximal subsets of  $A$  which meet matchings, say  $N_1$  and  $N_2$ , respectively. Consider the subgraph

\*This paper is the third in a series [1, 2]. It is, however, self-contained. W. A. of the first author is supported by the Army Research Office Durham through the NBS Combinatorial Mathematics Project. The second author is at the RAND Corporation, Santa Monica, Calif. His work is sponsored by the U.S. Air Force Project RAND. This paper was presented at the Advanced Study Institute on Integer Programming and Network Flow, Tahoe City, Calif., August 21 to July 1, 1965. S. I. A. Nash-Williams, prompted like ourselves by the same earlier papers, developed Theorems 1c, 2c, 1d, and 2d in another way. We are grateful for his correspondence, which has benefited our own work. We are indebted to Gian-Carlo Rota for his NBS Matroid Seminar lectures, which inspired the discovery of transversal matroids as well as a number of other ideas not yet set forth.

Figures in brackets indicate the literature references at the end of this paper.

$N \subset G$  formed by the edge-set

$$N_1 + N_2 = (N_1 - N_2) \cup (N_2 - N_1)$$

and the endpoints of its members. The connected components of  $N$  are simple open and closed paths because each node of  $N$  meets either one or two edges of  $N$ . Set

$$T_1 + T_2 = (T_1 - T_2) \cup (T_2 - T_1)$$

consists precisely of the path-ends of  $N$  that are in  $A$ ;  $(T_1 - T_2)$  are the nodes of  $A$  that meet  $N_1$  but not  $N_2$ , and  $(T_2 - T_1)$  are the nodes of  $A$  that meet  $N_2$  but not  $N_1$ . Suppose  $T_2$  is larger than  $T_1$ ; then  $T_2 - T_1$  is larger than  $T_1 - T_2$ . In this case, some component of  $N$  must be an open path, say  $P$ , which has one end  $v$  in  $T_2 - T_1$  and the other end not in  $T_1 - T_2$ . Regarding path  $P$  as its edge-set,  $N_1 + P = (N_1 - P) \cup (P - N_1)$  is a matching. This matching meets  $T_1$  in  $A$  and in addition it meets  $v$  in  $A$ . Thus, we contradict the hypothesis that  $T_1$  is a maximal subset of  $A$  which meets a matching. Therefore,  $T_1$  and  $T_2$  have the same cardinality and it follows that  $M_{G,E} = (E, F)$  is a matroid.

General matching matroids are discussed in section 6.

## 2. Introduction

P. J. Higgins [4] gives conditions for a family  $Q$  of sets to have  $k$  mutually disjoint partial transversals of prescribed sizes  $n_1, n_2, \dots, n_k$ . In section 4 we present conditions for a matroid  $M$  to have  $k$  mutually disjoint independent sets of prescribed sizes  $n_1, n_2, \dots, n_k$ . Where the matroid is graphic, for example, this result is new.

The following two closely related matroid theorems are presented in [1] and [2] as generalizations of theorems by Nash-Williams and Tutte on graphs. Theorem 2, below, for the case of transversals, handles a special case of the Higgins problem; it will be generalized to cover the Higgins problem. Theorem 1 is new for the case of transversals; it will be generalized analogously.

**THEOREM 1.** *The elements  $E$  of a matroid  $M$  can be partitioned into as few as  $k$  sets, each independent in  $M$ , if and only if  $|A| \leq k \cdot r(A)$  for all  $A \subset E$ .*

**THEOREM 2.** *The elements  $E$  of a matroid  $M$  can be partitioned into as many as  $k$  sets, each a spanning set of  $M$ , if and only if  $|A| \geq k(r(E) - r(\bar{A}))$  for all  $A \subset E$ .*

As usual  $|A|$  denotes cardinality of set  $A$ , and  $\bar{A}$  denotes the complement of  $A$  (with respect to  $E$ ). A spanning set of a matroid  $M$  is a subset of  $E$  which contains a maximal independent set.

A base of a matroid  $M$  is a maximal independent set, i.e., a minimal spanning set. Each base has cardinality equal to  $r(E)$ , the rank of the matroid.

For any family  $B$  of subsets of a set  $E$ , a covering in  $B$  is a subfamily whose union is  $E$ , and a packing in  $B$  is a subfamily whose members are disjoint.

Where  $B$  is the family of bases of matroid  $M$ , theorem 1 describes the minimum cardinality of a covering in  $B$ , and theorem 2 describes the maximum cardinality of a packing in  $B$ .

Applied to a transversal matroid  $M_b$ , where the members of a family  $Q$  are the matroid elements and where the subfamilies that have transversals are the independent sets of elements. Theorem 1 says that a family  $Q$  of sets can be partitioned into as few as  $k$  subfamilies, each having a transversal, if and only if  $|A| \leq k \cdot \rho(A)$  for every subfamily  $A \subset Q$ . Here  $\rho(A)$  denotes the maximum cardinality of a subfamily of  $A$  which has a transversal, i.e., the maximum cardinality of a partial transversal of  $A$ . The statement is not interesting when  $k=1$ ; for abstract matroids there is nothing interesting to say in this case.

Where  $A$  is a family of subsets of a set  $E$ , where  $N(E, A)$  is the  $(0, 1)$ -incidence matrix of members of  $E$  (rows) versus members of  $A$  (columns), and where  $G(E, A)$  is the bipartite incidence graph of  $(E, A)$ , the value  $\rho(A)$  is called the term rank of  $A$ ,  $N(E, A)$ , and  $G(E, A)$ , respectively. One of the two fundamental forms of the fundamental theorem of transversal theory is due to P. Hall. It describes when a family  $A$  (or  $Q$ ) itself has a transversal. The other fundamental form of the fundamental theorem is König's formula for term rank:  $\rho(A)$ , the maximum cardinality of a partial transversal of  $A$  or of a matching in  $G(E, A)$  (i.e., a set of 1's which might be called a matching in  $N(E, A)$ ) is equal to the minimum cardinality of a set of nodes that meets all edges in  $G(E, A)$  (i.e., a set of rows and columns that together contain all 1's of  $N(E, A)$ ).

Let  $\sigma(A)$ , for  $A \subset Q$ , denote the cardinality of the union of the members of  $A$ . It is a consequence of the König formula for term rank that the inequalities  $|A| \leq k \cdot \rho(A)$  for all  $A \subset Q$  are equivalent to the inequalities  $|A| \leq k \cdot \sigma(A)$  for all  $A \subset Q$ . Thus the latter are also necessary and sufficient for  $Q$  to have a partition into  $k$  subfamilies, each with a transversal. When  $k=1$ , this is P. Hall's theorem on transversals.

To see this equivalence, suppose that  $|A| > k \cdot \rho(A)$  for some  $A \subset Q$ . In the incidence graph  $G(E, A)$ , let  $E_1 \cup A_1$ ,  $E_1 \subset E$  and  $A_1 \subset A$ , be a minimum cardinality set of nodes that meets all of the edges. By the König theorem,  $\rho(A) = |E_1| + |A_1|$ . Let  $A_2 = A - A_1$ . The set-union of members of  $A_2$ , that is, the other ends of all the edges that meet  $A_2$ , is  $E_1$ , so  $\sigma(A_2) = |E_1|$ . Combining, we have

$$\begin{aligned} |A_2| &= |A| - |A_1| > k(|E_1| + |A_1|) - |A_1| \\ &= k \cdot |E_1| + (k-1) \cdot |A_1| \geq k \cdot \sigma(A_2). \end{aligned}$$

On the other hand, clearly  $\rho(A) \leq \sigma(A)$  for all  $A \subset Q$ . Therefore,  $|A| \leq k \cdot \rho(A)$  for all  $A \subset Q$  is equivalent to  $|A| \leq k \cdot \sigma(A)$  for all  $A \subset Q$ . Thus,  $Q$  can be partitioned into as few as  $k$  subfamilies, each with a transversal, if and only if the latter holds.

We do not recommend this matroid approach as the way to derive the transversal result. Theorem 1 in

general is not easy, and, even after it is established, using it with the König theorem to get the transversal result is no easier than deriving the transversal result directly from P. Hall's theorem as follows. Let each element  $e \in E$  be replicated  $k$  times to obtain  $e_1, \dots, e_k \in E'$ . To obtain  $Q'$ , let  $q' \in Q'$  consist of all the replications of the elements in  $q \in Q$ . Then  $|A| \leq k \cdot \sigma(A)$  for all  $A \subset Q$  is equivalent to  $|A'| \leq \sigma(A')$  for all  $A' \subset Q'$ . By P. Hall's theorem the latter is equivalent to the existence of a transversal for  $Q'$ . That, in turn, is equivalent to there being a partition of  $Q$  into as few as  $k$  subfamilies, each having a transversal.

Section 3 presents a derivation of transversal theorems using network flows. Section 4 presents a different derivation of the corresponding matroid theorems. Both derivations suggest computationally good algorithms. Section 5 presents another application of section 4, and section 6 relates general matching matroids to transversal matroids.

### 3. Transversal Covers and Packings

In this section we focus attention on the transversal matroid  $M_n = (E, F)$ ,  $F$  being the family of partial transversals of  $Q$ . We shall use network flows to derive results on covers and packings in  $F$ . For background material on network flows, we refer to [3]. In particular, the max-flow min-cut theorem and integrity theorem will be applied.<sup>3</sup>

Consider the directed network shown in figure 1. In figure 1 we have, in addition to a source-node  $u$  and a sink-node  $v$ , three tiers of nodes:  $e_1, e_2, \dots, e_n$  (elements of  $E$ );  $q_1, q_2, \dots, q_m$  (subsets of  $E$  that comprise the family  $Q$ ); and  $p_1, p_2, \dots, p_k$  (partial transversals). The directed edges of this network and their flow capacities are listed below:

	Edges	Capacities
$(u, e_j)$	$j = 1, \dots, n,$	$c(u, e_j) = 1,$
$(e_j, q_i)$	$e_j$ corresponding to $e_j \in q_i,$	$c(e_j, q_i) = \infty,$
$(q_i, p_r)$	$i = 1, \dots, m; \quad r = 1, \dots, k,$	$c(q_i, p_r) = 1,$
$(p_r, v)$	$r = 1, \dots, k,$	$c(p_r, v) = n_r.$

An integral flow from source to sink in this network produces  $k$  mutually disjoint partial transversals of respective sizes  $s_1 \leq n_1, s_2 \leq n_2, \dots, s_k \leq n_k$  in the following manner. Take a chain decomposition of the flow and put  $e_j$  in  $p_r$  if, for some  $i = 1, 2, \dots, m$ , the edges  $(e_j, q_i)$  and  $(q_i, p_r)$  occur in a chain of this decomposition. Conversely,  $k$  mutually disjoint partial transversals of sizes  $s_1 \leq n_1, s_2 \leq n_2, \dots, s_k \leq n_k$  yield an integral flow from source to sink. Using the

<sup>3</sup>In a graph where the edges  $e_i$  are directed and have positive integer capacities  $c_i$ , the maximum number of chains (directed paths, not necessarily distinct) from a node  $u$  to a node  $v$ , such that each  $e_i$  is contained in at most  $c_i$  of these chains, equals the minimum of the total capacity of the edges directed from  $u$  to  $v$  where  $(l, l')$  is any partition of all the nodes into two parts such that  $u \in l$  and  $v \in l'$ . The family of chains is called a chain decomposition of a maximum flow from  $u$  to  $v$ . The set of edges directed from  $l$  to  $l'$  is called a cut separating source  $u$  from sink  $v$ .

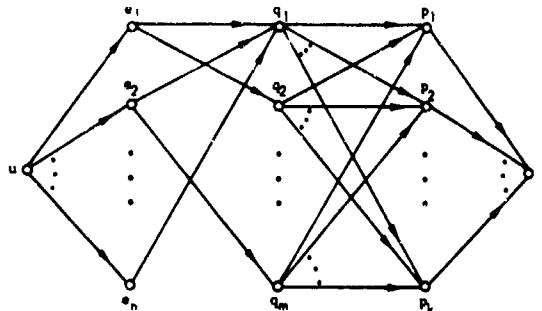


FIGURE 1.

integrity theorem and max-flow min-cut theorem for network flows, it follows that the maximum number of partial transversals of respective sizes  $s_1 \leq n_1, s_2 \leq n_2, \dots, s_k \leq n_k$  is equal to the capacity of a minimum cut separating source and sink in this network. We proceed to calculate this.

Let  $A, B, C$  be arbitrary subsets of  $E = \{e_1, e_2, \dots, e_n\}$ ,  $Q = \{q_1, q_2, \dots, q_m\}$ , and  $P = \{p_1, p_2, \dots, p_k\}$ , respectively, and denote their complements in these sets by  $\bar{A}, \bar{B}, \bar{C}$ . The capacity of an arbitrary cut separating  $u$  and  $v$  is then represented by the sum

$$\sum_{e_j \in A} c(u, e_j) + \sum_{\substack{e_j \in A \\ q_i \in B}} c(e_j, q_i) + \sum_{\substack{q_i \in B \\ p_r \in C}} c(q_i, p_r) + \sum_{p_r \in C} c(p_r, v).$$

We wish to minimize this over  $A \subset E, B \subset Q, C \subset P$ . Using the table of edge capacities, this reduces to computing the minimum of

$$|\bar{A}| + |\bar{B}| \cdot |\bar{C}| + \sum_{p_r \in C} n_r$$

over  $A \subset E, B \subset Q, C \subset P$  such that the set of edges leading from  $A$  to  $B$  is empty. Thus, for given  $A$  and  $C$ , we may take  $B$  to consist solely of those nodes of  $Q$  which are joined by edges to some node of  $A$ . In the language of set representatives,  $B$  consists of those sets represented by elements of  $A$ . Moreover, for  $\bar{C}$  of fixed cardinality  $|\bar{C}| = k - s$ , we may take  $C$  to correspond to the  $s$  smallest  $n$ 's. Thus, choosing the notation so that  $0 < n_1 \leq n_2 \leq \dots \leq n_k$ , and letting  $\sigma(A)$  denote the cardinality of the subfamily of  $Q$  represented by elements of  $A$ , we are led to minimizing

$$|\bar{A}| + (k - s) \sigma(A) + \sum_{r=1}^s n_r$$

over  $A \subset E$  and  $s = 0, 1, \dots, k$ . For fixed  $A$ , the minimization over  $s$  can be carried out explicitly. Indeed, let  $n_j^*$  be the number of integers among the  $n_r$ ,  $r = 1, 2, \dots, k$ , such that  $n_r \geq j$ ,  $j = 1, 2, \dots$ . Thus  $\{n_j^*\}$  and  $\{n_r\}$  are conjugate partitions of the integer  $\sum_{r=1}^k n_r$ . It is not hard to see, especially in terms of a

partition diagram, that

$$\min_{0 \leq s \leq k} \left[ (k-s)\sigma(A) + \sum_{r=1}^s n_r \right] = \sum_{j=1}^{\sigma(A)} n_j^*.$$

This proves that the maximum number of elements of  $E$  contained in a union of  $k$  (mutually disjoint) partial transversals of  $Q$ , having respective sizes  $s_1 \leq n_1, s_2 \leq n_2, \dots, s_k \leq n_k$ , is equal to

$$(*) \quad \min_{A \in E} \left[ |\bar{A}| + \sum_{j=1}^{\sigma(A)} n_j^* \right].$$

Here  $\sigma(A)$  denotes the number of sets in the family  $Q$  that are represented by elements of  $A$ .

The following two theorems, which give necessary and sufficient conditions for the existence of covers and packings composed of partial transversals of prescribed sizes, are consequences of this result. (Nash-Williams originated a similar viewpoint for related theorems on matroids.)

**THEOREM 1a.** Let  $Q$  be a finite family of subsets of a finite set  $E$ . The family  $Q$  has  $k$  partial transversals of respective sizes  $n_1, n_2, \dots, n_k$  whose union is  $E$  if and only if (i)  $n_i \leq \rho(E)$ ,  $i=1, 2, \dots, k$ , and (ii) for every  $A \subset E$ , the inequality

$$|A| \leq \sum_{j=1}^{\sigma(A)} n_j^*$$

holds.

Here  $\rho(E)$  denotes the term rank of the bipartite incidence graph (or matrix) of elements of  $E$  versus sets of the family  $Q$ , that is,  $\rho(E)$  is the rank of the matroid  $M_Q = (E, F)$ . The proof of sufficiency of (i) and (ii) makes use of the fact that  $M_Q$  is a matroid in extending the  $k$  partial transversals of sizes  $s_i$  to partial transversals of sizes  $n_i$ ,  $i=1, 2, \dots, k$ .

**THEOREM 2a.** Let  $Q$  be a finite family of subsets of a finite set  $E$ . The family  $Q$  has  $k$  mutually disjoint partial transversals of respective sizes  $n_1, n_2, \dots, n_k$  if and only if, for every  $A \subset E$ , the inequality

$$|A| \geq \sum_{j=\sigma(A)+1}^k n_j^*$$

holds.

Using the König theorem in an argument similar to that in section 2 shows that the rank function  $\rho(A)$  of matroid  $M_Q$  can be used in place of  $\sigma(A)$  in (\*), hence also in theorems 1a and 2a.

The situation of theorem 2a is the problem studied by Higgins. His conditions are not the same as those of theorem 2a, but are instead stated in terms of subfamilies  $B$  of  $Q$  rather than subsets  $A$  of  $E$ . They may be derived from theorem 2a by use of the König theorem (and vice versa), or can be obtained directly by eliminating  $A$  and  $C$ , rather than  $B$  and  $C$ , in the minimization argument leading to (\*).

#### 4. Matroid Partition

**THEOREM 1b.** The set  $E$  of elements of a matroid  $M$  can be covered by a family of independent subsets  $I_i$  ( $i=1, \dots, k$ ) of prescribed sizes  $n_i \leq r(E)$  if and only if, for every  $A \subset E$ ,

$$|A| \leq \sum_{j=1}^{\sigma(A)} n_j^* = \sum_i \min(n_i, r(A)).$$

**THEOREM 2b.** The set  $E$  of elements of a matroid  $M$  contains mutually disjoint independent subsets  $I_i$  ( $i=1, \dots, k$ ) of prescribed sizes  $n_i \leq r(E)$  if and only if, for every  $A \subset E$ ,

$$|A| \geq \sum_{j=\sigma(A)+1}^{\sigma(E)} n_j^* = \sum_i [n_i - \min(n_i, r(\bar{A}))].$$

Here  $r(A)$  denotes rank relative to matroid  $M$ . The equations in theorems 1b and 2b are obvious.

Using lemma 1, theorems 1b and 2b follow immediately from theorems 1c and 2c below.

**LEMMA 1.** For any matroid  $M=(E, F)$  and any non-negative integer  $n$ , let  $F_{(n)}$  denote the members of  $F$  which have cardinality at most  $n$ . Then  $M_{(n)}=(E, F_{(n)})$  is a matroid. Where  $r(A)$  is the rank function for  $M$ , the rank function for  $M_{(n)}$  is

$$r_{(n)}(A) = \min(n, r(A)).$$

We call  $M_{(n)}$  the truncation of  $M$  at  $n$ .

The proof of lemma 1 is obvious.

Let  $r_i(A)$  be the rank functions for any family of matroids  $M_i=(E, F_i)$ ,  $i=1, \dots, k$ , on the set  $E$  of elements.

**THEOREM 1c.** Set  $E$  can be partitioned into a family of subsets  $I_i$  ( $i=1, \dots, k$ ), where  $I_i \in F_i$ , if and only if for every  $A \subset E$ ,

$$|A| \leq \sum_i r_i(A).$$

**THEOREM 2c.** There is a family of mutually disjoint sets  $I_i$  ( $i=1, \dots, k$ ), where  $I_i$  is a maximal member (base) in  $F_i$ , if and only if for all  $A \subset E$ ,

$$|A| \geq \sum_i r_i(E) - \sum_i r_i(\bar{A}).$$

Where each  $M_i$  is a graph, theorem 2c is equivalent to a theorem of Tutte [5].

Since it can be shown that a truncation of a graphic or a transversal matroid is not necessarily graphic or transversal, theorems 1b and 2b for these cases do not follow from theorems 1c and 2c for these cases as in general. A similar remark applies to the way 2c will be derived from 1c. Thus we observe that the general matroid concept is useful even where primary interest is more special. The proof of 1c, on the other hand, is arranged so that the only matroids it will mention are those of the theorem. Hence, the proof

applies directly to any special class of matroids (including classes of one). Everything in references [1] and [2] applies directly to the case of only graphs. The proofs in [2] do not apply directly to the case of only transversals because, as will be shown at another time, a "contraction" of a transversal matroid is not necessarily transversal.

**LEMMA 2.** Let  $A$  be any subset of the elements of a matroid  $M$ . Let  $I$  be any independent subset of  $A$ . A maximal set  $S$ , such that  $I \subset S \subset A$  and  $r(S) = r(I) = |I|$ , is the unique set consisting of  $I$  and elements  $e \in A$  such that  $e \cup I$  is dependent.

Set  $S$  is called the span of  $I$  in  $A$ .

**PROOF.** Consider  $e \in A - I$ . By the definition of rank,  $I$  is a maximal independent subset of any  $S$ . Thus, if  $e \cup I$  is independent, then  $e \in S$ . And thus, on the other hand, if  $e \cup I$  is dependent, then  $I$  is a maximal independent subset of  $e \cup S$ . Hence by axiom 2 for matroids,  $r(e \cup S) = |I|$ , and so  $e \in S$ .

**LEMMA 3.** The union of any independent set  $I$  and any element  $e$  of a matroid  $M$  contains at most one minimal dependent set.

A minimal dependent set is called a circuit of  $M$ .

**PROOF.** Suppose  $I \cup e$  contains two distinct circuits  $C_1$  and  $C_2$ . Assume  $I$  is minimal for this possibility. We have  $e \in C_1 \cap C_2$ . There is an element  $e_1 \in C_1 - C_2$  and an element  $e_2 \in C_2 - C_1$ . Set  $(I \cup e) - (e_1 \cup e_2)$  is independent since otherwise  $I - e_1$  is a smaller independent set than  $I$  for which  $(I - e_1) \cup e$  contains more than one circuit. Set  $I$  and set  $(I \cup e) - (e_1 \cup e_2)$  are maximal independent subsets of set  $I \cup e$ . This contradicts axiom 2.

**PROOF OF 1c.** Suppose that  $\{I_i\}$  ( $i = 1, \dots, k$ ) is a partition of  $E$ , where  $I_i \in F_i$ . Then for arbitrary  $A \subset E$ ,

$$|A| = \sum_i |A \cap I_i| = \sum_i r_i(A \cap I_i) \leq \sum_i r_i(A).$$

Conversely, suppose that for every  $A \subset E$ , the inequality holds. Let  $\{I_i\}$  ( $i = 1, \dots, k$ ) be a family of disjoint sets such that  $I_i$  is independent in  $M_i$ . Any number of these may be empty. Suppose there is an

$$e \in E - \bigcup_i I_i.$$

We shall show how to rearrange elements among the sets  $I_i$  to make room for  $e$  in one of them while preserving the mutual disjointness and the independence of  $I_i$  in  $M_i$ . This will prove the theorem.

If  $e \in S$  for any  $S \subset E$ , then for some  $i$ ,  $|I_i \cap S| < r_i(S)$ . Otherwise,

$$\begin{aligned} |S| &\geq \left| \bigcup_i (I_i \cap S) \cup e \right| \\ &= 1 + \sum_i |I_i \cap S| > \sum_i r_i(S) \end{aligned}$$

would contradict the hypothesis.

Let  $S_0 = E$ . Inductively, starting with  $j-1=0$ , if  $e \in S_{j-1}$  then for some  $I_{(j)}$  such that

$$|I_{(j)} \cap S_{j-1}| < r_{(j)}(S_{j-1}),$$

we define  $S_j$  to be the span in  $S_{j-1}$ , with respect to matroid  $M_{(j)}$ , of  $I_{(j)} \cap S_{j-1}$ . Since

$$r_{(j)}(S_j) < r_{(j)}(S_{j-1}),$$

$S_j$  is a proper subset of  $S_{j-1}$ . Therefore we must eventually reach an  $S_h$  such that  $e \notin S_h$  and  $e \in S_j$  for  $0 \leq j < h$ .

(Where the matroids  $M_i$  are identical, the construction above is the same as the corresponding part of the proof of theorem 1 in [1]. The rest of references [1] and [2] goes through essentially unchanged for a version, concerning possibly distinct matroids, which includes theorems 1c and 2c. However, we continue here with a substantially trimmed version.)

If  $e \cup I_{(h)}$  is independent in  $M_{(h)}$ , the present proof is finished. Otherwise  $e \cup I_{(h)}$  contains a circuit  $C$  of  $M_{(h)}$ . Set  $(e \cup I_{(h)}) \cap S_{h-1}$  is not dependent in  $M_{(h)}$ , because then, by lemma 2 and by the definition of  $S_h$ , since  $e \in S_{h-1}$ , we would have  $e \in S_h$ . Thus let  $m$  be the smallest integer,  $0 < m < h$ , such that  $(e \cup I_{(h)}) \cap S_m$  is independent in  $M_{(h)}$ . There is an  $e' \in C - S_m$ . By lemma 3,  $e \cup I_{(h)} - e'$  is independent in  $M_{(h)}$ .

Replacing  $I_{(h)}$  by  $e \cup I_{(h)} - e'$ , we now need to dispose of  $e'$  instead of  $e$ . However, we can show that sequence  $(I_{(1)}, S_1), \dots, (I_{(m)}, S_m)$ , with the roles of  $e$  and  $e'$  interchanged, is of the same construction as  $(I_{(1)}, S_1), \dots, (I_{(h)}, S_h)$ , only shorter. Since the original  $e \cup (I_{(h)} \cap S_{j-1})$  is dependent in  $M_{(h)}$ , for all  $j$ ,  $1 \leq j \leq m$ , by lemma 3 we have  $e' \in C \subset S_{j-1}$ . Consider the terms  $(I_{(j)}, S_j)$ ,  $1 \leq j \leq m$ , one after another in order. Assume there is no change in  $S_{j-1}$ . If originally  $I_{(j)} \neq I_{(h)}$ , then there is no change at all in  $(I_{(j)}, S_j)$ . If originally  $I_{(j)} = I_{(h)}$ , then, even though  $e$  and  $e'$  are interchanged in  $I_{(j)}$ , by lemma 2 and the definition of  $S_j$ , since  $e \cup e' \subset C \subset S_{j-1}$ , there is no change in  $S_j$ . Thus the theorem is proved.

**PROOF OF 2c.** For any family of matroids,  $M_i = (E, F_i)$  ( $i = 1, \dots, k$ ), with rank functions  $r_i(A)$ , consider the additional matroid  $M_0 = (E, F_0)$  where the members of  $F_0$  are the subsets of  $E$  that have car-

dinality at most  $|E| - \sum_i r_i(E)$ . Matroid  $M_0$  is a trun-

cation of the matroid in which all subsets of  $E$  are independent. The existence of mutually disjoint sets  $I_i$  ( $i = 1, \dots, k$ ), where  $I_i$  is a maximal member of  $F_i$ , is equivalent to the existence of a partition of  $E$  into a family of sets,  $I_0$  and  $I_i$  ( $i = 1, \dots, k$ ), such that  $I_0 \in F_0$  and  $I_i \in F_i$ . By theorem 1c, the existence of that partition is equivalent to the condition that

$$|A| \leq \min(|E| - \sum_i r_i(E), |A|) + \sum_i r_i(A)$$

for all  $A \subset E$ .

That condition in turn is equivalent to

$$|A| \leq |E| - \sum_i r_i(E) + \sum_i r_i(A)$$

for all  $A \subset E$ , which is equivalent to

$$|A| \geq \sum_i r_i(E) - \sum_i r_i(A)$$

for all  $A \subset E$ . Thus theorem 2c is proved.

### 5. Another Application

Let  $J_i$  ( $i=1, \dots, k$ ) be mutually disjoint independent sets in a matroid  $M=(E, F)$ . Let  $E' = E - (\cup J_i)$ .

**THEOREM 1d.** Set  $E$  can be partitioned into a family of independent sets  $I_i \in F$  ( $i=1, \dots, k$ ) such that  $J_i \subset I_i$  if and only if, for every  $A \subset E'$ ,

$$|A| \leq \sum_i [\pi(A \cup J_i) - \pi(J_i)].$$

**THEOREM 2d.** There is a family of mutually disjoint bases  $I_i$  ( $i=1, \dots, k$ ) of  $M$  such that  $J_i \subset I_i$  if and only if, for every  $A \subset E'$ ,

$$|A| \geq \sum_i [\pi(E) - \pi(E' - A \cup J_i)].$$

For any matroid  $M=(E, F)$  and any  $E_0 \subset E$ , let  $F_0$  consist of sets  $I \in F$  such that  $I \subset E_0$ . Then  $M \cdot E_0 = (E_0, F_0)$ , obviously a matroid, is called a *submatroid* of  $M$  (obtained from  $M$  by deleting the elements of  $\bar{E}_0 = E - E_0$ ). The rank of any subset of  $E_0$  is the same in  $M \cdot E_0$  as in  $M$ .

For any matroid  $M=(E, F)$  and any  $E_0 \subset E$ , let  $J$  be any maximal subset of  $\bar{E}_0 = E - E_0$  which is a member of  $F$ . In other words, let  $J$  be any base of submatroid  $M \cdot \bar{E}_0$ . Let  $F_0$  consist of sets  $I \in F$  such that  $I \subset E_0$  and such that  $J \cup I \in F$ . It follows easily from the definition of matroid that  $M \times E_0 = (E_0, F_0)$  is a unique matroid, called the *contraction* of  $M$  to  $E_0$  (obtained from  $M$  by contracting the elements of  $\bar{E}_0$ ). Where  $r$  and  $r_0$  denote the rank functions for matroids  $M$  and  $M \times E_0$ , respectively, we have for every  $A \subset E_0$ ,

$$r_0(A) = \pi(A \cup \bar{E}_0) - \pi(\bar{E}_0).$$

Theorem 1d follows immediately from theorem 1c by letting the  $M_i$  of 1c (for  $i=1, \dots, k$ ) be the matroid obtained from matroid  $M$  of 1d by contracting the elements of  $J_i$  and then deleting all the other elements of  $E - E'$ .

To prove 2d from 2c, we obtain each  $M_i$  of 2c from  $M$  of 2d in the same way as above. If, for some  $i$ ,  $\pi(E' \cup J_i) < \pi(E)$ , then no base of  $M$  is contained in  $E' \cup J_i$  and so there is no family of bases  $I_i$  as described in 2d. In this case the inequality in 2d does not hold where  $A$  is the empty set. Otherwise,  $\pi(E' \cup J_i) = \pi(E)$  for each  $i$ . In this case, if  $J'_i$  is a base of  $M_i$ , then  $J_i \cup J'_i$  is a base of  $M$ . Thus, in this case, 2d follows from 2c.

### 6. Addendum on Matchings

An element of a matroid  $M$  is called *isolated* if it is contained in every base of  $M$ , i.e., if it is contained in no circuits of  $M$ . Clearly, any number of isolated elements can be "added" to any transversal matroid  $M_a$ , thereby obtaining another transversal matroid. With respect to the graph representation  $G(E, Q)$  of  $M_a$ , for every isolated element  $e'$  added to  $M_a$ , simply add a node  $e'$  to  $E$  and join it to a new node  $q'$  added to  $Q$ .

Several elements of a matroid  $M$  are said to be *in series* with each other either when they are all isolated, or else when none of them is isolated and each base of  $M$  contains all but possibly one of them.

A set of elements is *in series* in matroid  $M$  if and only if the elements are contained in exactly the same circuits of  $M$ .

Suppose some base  $I$  of  $M$  contains neither of elements  $e_1$  and  $e_2$  of  $M$ . Then  $I \cup e_1$  contains a circuit of  $M$  that contains  $e_1$  but not  $e_2$ .

Suppose an element  $e_1$  is contained in a circuit  $C$  of  $M$  that does not contain nonisolated element  $e_2$  of  $M$ . Let  $I$  be a base of  $M$  which does not contain  $e_2$ . The rank of  $(I \cup C) - e_1$  is as large as the rank of  $I$ ; otherwise every maximal independent subset of  $I \cup C$  would contain  $e_1$ , but then  $e_1$  would be contained in no circuit in  $I \cup C$ . Therefore  $(I \cup C) - e_1$  contains a base of  $M$ ; this base contains neither  $e_1$  nor  $e_2$ . Thus the theorem is proved.

"Replacing an element  $e_i$  in a matroid  $M$  by a set  $E_i^* = \{e_i^1, \dots, e_i^k\}$  of new elements in series" yields a matroid  $M^{(i,k)}$ . The circuits of  $M^{(i,k)}$  and the elements of  $M^{(i,k)}$  are identical with those of  $M$  except that  $e_i$  is replaced by the members of  $E_i^*$ . Each base  $B$  of  $M$  which contains  $e_i$  corresponds to a base  $(B - e_i) \cup E_i^*$  of  $M^{(i,k)}$ . Each base  $B$  of  $M$  which does not contain  $e_i$  corresponds to  $k$  bases of  $M^{(i,k)}$  of the form  $B \cup E_i^* - e_i^j$ ,  $j=1, \dots, k$ . We omit proof that  $M^{(i,k)}$  is a matroid, which is not difficult using the description of the bases.

For any transversal matroid  $M_a$ , containing element  $e_i$ , the matroid  $M_a^{(i,k)}$  is also transversal.

Let  $M_a$  be represented by a bipartite graph  $G = G(E, Q)$  as described in section 1; a base of  $M_a$  consists of the endpoints in  $E$  of a maximum cardinality matching in  $G$ . By thinking of bases, it is easy to see that we obtain from  $G$  a similar representation  $G^{(i,k)}$  for matroid  $M_a^{(i,k)}$  as follows. Replace node  $e_i \in E$  of  $G$  by the set  $E_i^*$  of new nodes. Join each  $e_j \in E_i^*$  to the same nodes in  $Q$  to which  $e_i$  was joined. Also add to  $G$  a set  $Q'$  of  $k-1$  new nodes, each joined to precisely the members of  $E_i^*$ . We then have  $G^{(i,k)}$ . A base of matroid  $M_a^{(i,k)}$  consists of the endpoints in  $(E - e_i) \cup E_i^*$  of a maximum cardinality matching in  $G^{(i,k)}$ .

Clearly, if  $A \subset E$  for matching matroids  $M_{G,A}$  and  $M_{G,K}$ , then  $M_{G,A}$  is the submatroid of  $M_{G,K}$  whose set of elements is  $A$ . Clearly, any submatroid of a transversal matroid is transversal.

Every matching matroid is a transversal matroid. (Thus, the two classes of matroids are abstractly the same.)

In view of the preceding observations on submatroids, it suffices to show that where  $G$  is any graph and where  $V$  is all of its nodes,  $M_{G,V}$  is a transversal matroid. Clearly,  $B$  is a base in matroid  $M_{G,V}$  if and only if  $B$  is the set of endpoints of some maximum (cardinality) matching  $L$  in  $G$ .

Section 6 of [7] implies the following theorem (which essentially strengthens some other known theorems, a characterization by Tutte of graphs in which no matching meets all the nodes, and a formula by Berge for what we regard here as the rank of  $M_{G,V}$ ).

(\*) From any graph  $G$ , by deleting the set  $J$  of nodes which meet every maximum (cardinality) matching and deleting all the edges which meet  $J$ , the remainder consists of connected components,  $O_i$ , containing respectively  $2r_i + 1$  nodes where  $r_i$  is an integer. (If  $G$  is bipartite, each  $O_i$  is a single node.) Let  $Q$  consist of the nodes  $u$  in  $J$  which in  $G$  are joined to at least one node in  $\cup O_i$ . Every maximum matching in  $G$  contains  $r_i$  edges in  $O_i$ , for each  $i$ , and contains an edge joining  $u$  to a node in  $\cup O_i$ , for each  $u \in Q$ .

What is actually proved in [7] is theorem (\*) where "Every" is replaced by "Some" in the last sentence. However, because each  $O_i$  has an odd number of nodes, because every edge leaving an  $O_i$  goes to a  $u \in Q$ , and because each edge has two ends, it is easy to see that any matching which is not as described in the theorem meets fewer nodes in  $\cup O_i$ . Hence, it has smaller cardinality than the matching, described in the theorem, which is proved in [7] to exist.

(Unless some matching in  $G$  meets every node, there are more  $O_i$ 's than there are  $u$ 's. The theorem of Tutte says that a graph contains no matching that meets all of the nodes if and only if there exists a subset  $Q$  of the nodes such that deleting  $Q$  and its incident edges from  $G$  leaves more than  $|Q|$  components which have odd numbers of nodes.)

For any graph  $G$ , whose node set is  $V$ , the set  $J \subset V$  defined in (\*) is the set of isolated elements in matroid  $M_{G,V}$ . Denoting the set of nodes in  $O_i$  by  $E_i$ , theorem (\*) says that each maximum matching meets all but possibly one node in  $E_i$ ; thus, set  $E_i$  is in series in matroid  $M_{G,V}$ . By "contracting" the subgraphs  $O_i$  to single nodes  $e_i$ , comprising a set  $E$ , and then by deleting  $J - Q$  and all edges which do not meet an  $e_i$ , we obtain from  $G$  a bipartite graph  $G(E, Q)$ .

Let  $M_0$  be the transversal matroid, with set  $E$  of elements, associated with  $G(E, Q)$ . It follows easily from theorem (\*) that matroid  $M_{G,V}$  is obtained from matroid  $M_0$  by replacing each  $e_i$  by the set  $E_i$  in series and by adding set  $J$  of isolated elements.

The structure of transversal matroids and some other related matroids will be further described in a later paper.

## 7. References

- [1] Edmonds, J., Minimum Partition of a Matroid into Independent Subsets, J. Res. NBS 69B (Math. and Math. Phys.), Nos. 1 and 2, 67-72 (1965).
- [2] Edmonds, J., On Lehman's Switching Game and a Theorem of Tutte and Nash-Williams, J. Res. NBS 69B (Math. and Math. Phys.), Nos. 1 and 2, 73-77 (1965).
- [3] Ford, L. R., Jr., and D. R. Fulkerson, Flows in Networks (Princeton University Press, Princeton, N.J., 1962).
- [4] Higgins, P. J., Disjoint Transversals of Subsets, Can. J. Math. 11, 280-285 (1959).
- [5] Tutte, W. T., On the Problem of Decomposing a Graph into  $n$  Connected Factors, J. London Math. Soc., 36, 221-230 (1961).
- [6] Tutte, W. T., Lectures on Matroids, J. Res. NBS 69B (Math. and Math. Phys.), Nos. 1 and 2, 1-47 (1965).
- [7] Edmonds, J., Paths, Trees and Flowers, Can. J. Math., 17, No. 3, 449-467 (1965).

(Paper 69B3-145)

Note added in proof: Theorem 1, the subject of [1], generalized here, was proved for the case where the matroid is a set of vectors in a vector space by Alfred Horn [A characterization of unions of linearly independent sets, J. London Math. Soc. 30 (1955), 494-496] and by R. Rado [A combinatorial theorem on vector spaces, J. London Math. Soc. 37, (1962), 351-353]. In the Abstracts of Short Communications, International Congress of Mathematicians, Stockholm 1962, p. 47, Rado remarks that "This theorem is of interest since in contrast to other propositions on vector spaces its proof has not yet been extended to abstract independence relations  $I$  (H. Whitney, Amer. J. Math. 1935, R. Rado, Canadian J. Math. 1949). It remains to decide if (i) the theorem is true for all  $I$ , or (ii) its validity constitutes a new necessary condition for representability of  $I$  in a vector space." Theorem 1 confirms (i).

**INTEGER PROGRAMMING**

by

**RALPH E. GOMORY**

at the

**American Mathematical Society Summer Seminar**

on the

**Mathematics of the Decision Sciences**

**Stanford University**

**July - August 1967**

## ON THE RELATION BETWEEN INTEGER AND NONINTEGER SOLUTIONS TO LINEAR PROGRAMS\*

By R. E. GOMORY

THOMAS J. WATSON RESEARCH CENTER, YORKTOWN HEIGHTS, NEW YORK

*Communicated by R. Courant, December 22, 1964*

We will refer to the ordinary linear programming problem

$$\begin{aligned} \text{maximize } z &= cx \\ Ax &= b, x \geq 0 \end{aligned} \tag{1}$$

as problem  $P1$ . In (1)  $b$  is an integer  $m$ -vector,  $c$  is an  $m + n$  vector, and  $A$  is an  $m \times (m + n)$  integer matrix.  $x$  is an  $m + n$  vector, all of whose components are required to be nonnegative. We assume that  $A$  is of the form  $(A', I)$  with  $I$  an  $m \times m$  identity matrix, so that in (1)  $Ax = b$  is equivalent to the  $m$  inequalities in  $n$  variables  $A'x' \leq b$ . We will say that  $x$  is feasible if it satisfies the equality and non-negativity conditions of (1) and optimal if it also maximizes.

A problem closely related to  $P1$  is the integer programming problem  $P2$  which is  $P1$  with the added condition that the components of  $x$  be integers. Because of the comparative ease with which  $P1$  is solved<sup>1</sup> and the comparative difficulty of  $P2$ ,<sup>2,3</sup> it is natural to consider getting from the solution of  $P1$  to the solution to  $P2$  by some sort of a "rounding" process through which the noninteger components of the  $x$  solving  $P1$  are rounded either up or down to produce a solution to  $P2$ . This

procedure seems particularly plausible when the components  $x_i$  of  $x$  are reasonably large numbers. However, it is easily shown by examples that a nearest-neighbor rounding process cannot generally produce the optimal solution to  $P2$ . These examples are neither pathological nor uncommon; it is simply not the case that the optimal solution can be obtained by simple rounding to some vector  $x'$  with  $|x'_i - x_i| < 1$ , even if the rounding is followed by some sort of optimization on the residual problem and even if the  $b$  and  $x$  of (1) become arbitrarily large.

Nevertheless, there is a close connection between the optimal solutions to  $P1$  and  $P2$  for a wide range of right-hand sides  $b$ . We first give some theorems on this connection and then an algorithm which for these  $b$  obtains the optimal solution of  $P2$  from the optimal solution to  $P1$ .

If  $B$  is a basis, i.e., an  $m \times m$  nonsingular submatrix of  $A$ , we will assume that  $A$  has been rearranged and partitioned into matrices  $B$  and  $N$  with  $A = (B, N)$ . We will also partition  $x = (x_B, x_N)$  and  $c = (c_B, c_N)$ . The columns of  $A$  will be referred to as  $\alpha_i$ ,  $B = (\alpha_1, \dots, \alpha_m)$ . We confine ourselves to right-hand side vectors  $b$  in that part of  $m$ -space for which (1) is solvable. If  $B$  is the optimal basis for  $P1$  with right-hand side  $b$ , then it is also the optimal basis for all  $b'$  such that  $B^{-1}b' \geq 0$ . These  $b$  form a cone in  $m$ -space, and in fact all solvable  $m$ -space is partitioned into such cones  $K^B$ . On removing from  $K^B$  all points within a distance  $d$  of its boundary, we have the reduced cone  $K^B(d)$ . With this notation we can now state Theorem 1.

**THEOREM 1.** Let  $l = \max \|\alpha_i\|$ ,  $i = m+1, \dots, m+n$ ,  $D = |\det B|$ , and  $z_1(b)$  be the value of the solution to  $P1$ . Then if  $b \in K^B(l(D-1))$ , the value  $z_2(b)$  of the solution to  $P2$  is given by

$$z_2(b) = z_1(b) + \varphi^B(b), \quad (2)$$

and an optimal solution vector is given by

$$x(b) = (x_B(b), x_N(b)) = (B^{-1}(b - Ny^B(b)), y^B(b)), \quad (3)$$

where both the scalar function  $\varphi^B(b)$  and the  $n$ -vector function  $y^B(b)$  are  $m$ -periodic, i.e.,  $\varphi^B(b + \alpha_i) = \varphi^B(b)$ ,  $i = 1, \dots, m$ , and  $y^B(b + \alpha_i) = y^B(b)$ ,  $i = 1, \dots, m$ .

The periodicity means that the values of  $\varphi^B(b)$  and  $y^B(b)$  depend only on the position of  $b$  relative to the lattice  $\mathcal{L}_B$  of points generated by integer combinations of  $\alpha_1, \dots, \alpha_m$ . This is equivalent to saying that  $\varphi^B$  is a function on the factor module  $M(I)/M(B)$  where  $M(I)$  is the module of all integer points in  $m$ -space and  $M(B)$  the module of integer combinations of the  $\alpha_i$ ,  $i = 1, \dots, m$ .

Although (2) and (3) have just the form one would expect if rounding were possible, the integer solution  $x(b)$  is generally not a continuation of a rounded nearest-neighbor solution. It is instead a continuation from a point  $p = b - Ny^B(b)$  which is on  $\mathcal{L}_B$ . A measure of the distance from  $p$  to  $b$  is given by Theorem 2.

**THEOREM 2.** If  $b \in K^B(l(D-1))$ , then the optimal solution vector  $x(b)$  has the property

$$\sum_{i=m+1}^{m+n} x_i = \sum_{i=1}^m y_i \leq D - 1.$$

We next discuss the arithmetic work involved in actually obtaining  $\varphi^B(b)$  and

$y^s(b)$ . The calculation may be broken into two parts, of which the first is a standard calculation.

The factor module  $M(I)/M(B)$  is a finite additive group having  $D$  elements. By the methods of references 4 or 5 we calculate  $M(I)/M(B)$  as the direct sum of cyclic groups of known orders. The arithmetic work involved is bounded by  $2m(m^2 + 2m)\log_2 D$ . Or, alternatively, given  $B^{-1}$  as a starting point, the standard form of  $M(I)/M(B)$  can be obtained in at most  $2r(m^2 + 2m)\log_2 D$  arithmetic steps, where  $r \leq m$  is the rank of  $M(I)/M(B)$ . If the factor group is cyclic,  $r$  is 1.

This calculation also provides explicitly a means of mapping integer  $m$ -vectors onto corresponding elements of  $M(I)/M(B)$ . If we call this mapping  $f$ , then  $\bar{p} = fp$  can be obtained for any integer  $m$ -vector  $p$  by at most  $m^2 + 2m$  arithmetic steps.

Since both  $\varphi^s(b)$  and  $y^s(b)$  are  $m$ -periodic, one can obtain, by periodicity, values  $\varphi^s(b)$  and  $y^s(b)$  for all  $b$  if they are known for one period.

**THEOREM 3.** *There are  $D$  distinct  $b$  values in one period. If  $M(I)/M(B)$  has been put in standard form and the group elements  $\bar{\alpha}_i = f\alpha_i$  obtained for  $i = m + 1, \dots, m + n$ , then the values of  $\varphi^s(b)$  can be computed for all  $b$  in one period in less than  $7nD$  elementary arithmetic steps. The values of  $y^s(b)$  for a particular  $b$  can be computed in  $n$  more steps or the values for all  $b$  in one period in  $nD$  more steps.*

The arithmetic steps referred to here are operations such as the addition and subtraction of real numbers, comparison of real numbers, or the addition and subtraction of elements of  $M(I)/M(B)$ . We now turn to the proofs of these theorems.

In reference 2 it was pointed out that  $M(I)/M(B)$  is isomorphic to the group  $F$  generated by the rows of the matrix  $B^{-1}A$  with the entries being replaced by the corresponding entries modulo 1. These "fractional rows" then provided the basic inequalities for the methods of reference 2 and, in a less evident way, for reference 3. As was remarked in reference 2, similar reasoning shows that  $M(I)/M(B)$  is also isomorphic to the group generated by the columns of  $B^{-1}A$  with coefficients being treated the same way, i.e., reduced to proper fractions. As  $B^{-1}A$  is the simplex tableau provided by the simplex method in solving (1), each column has associated with it a relative cost factor. This representation suggests the following problem involving maximization over  $M(I)/M(B)$ .

$$\max \sum_{i=1}^{i=n} c^*_{i+m} y_i$$

$$\sum_{i=1}^{i=n} \bar{\alpha}_{i+m} y_i = \bar{b}, y_i \geq 0 \text{ and integer.} \quad (4)$$

Here  $\bar{\alpha}_i$  and  $\bar{b}$  are the elements of  $M(I)/M(B)$  corresponding to the vectors  $\alpha_i$  and  $b$ , and  $c^*$ , is the relative cost  $c^*_i = c_i - c_B B^{-1} \alpha_i$ .

It is a fundamental property of linear programming that all  $c^*$ , associated with an optimal basis are  $\leq 0$ ; so the maximum in (4) does exist for optimal bases (though not for other bases).

Since the  $\bar{\alpha}_i \in M(I)/M(B)$ , a group with  $D$  elements,  $D\bar{\alpha}_i = \bar{0}$ ; so for a minimal solution to (4) it is only necessary to consider  $y_i$  satisfying  $0 \leq y_i < D$ . We will indicate later how (4) can, in fact, be solved for all  $\bar{b}$  in a total of  $7nD$  elementary steps.

If  $\bar{b}$  is an integer  $m$ -vector, define  $\varphi(b)$  as the value of the solution of (4) with  $\bar{b} = f\bar{b}$ . This is the  $\varphi^s$  of Theorem 1.

One of the properties of  $\varphi$  is immediate. Clearly,  $\varphi(b + \alpha_i) = \varphi(b)$ ,  $i = 1, \dots, m$ .

Also  $\varphi(b) + c_B B^{-1}b \geq z_1(b)$ . For if  $(x'_B, x'_N)$  is a feasible integer solution to (1), then the corresponding cost  $c(x'_B, x'_N) = c_B x'_B + c_N x'_N$  can be expressed in terms of the  $x'_N$  only by using the relation

$$Bx'_B + Nx'_N = b, \quad (5)$$

which yields

$$c(x'_B, x'_N) = c_B B^{-1}b + (c_N - c_B B^{-1}N)x'_N.$$

$c_B B^{-1}b$  is  $z_1(b)$ , so

$$c(x'_B, x'_N) = z_1(b) + \sum_{i=-m+1}^{i=-m+n} c^* x'_i. \quad (6)$$

Applying the homomorphism  $f$  to (5),  $B$  disappears and we get

$$\sum_{i=-m+1}^{i=-m+n} a_i x'_i = b.$$

so  $x'_N$  is a feasible solution to (4). Hence

$$\sum_{i=-m+1}^{i=-m+n} c^* x'_i \leq \varphi(b);$$

so from (6),  $c(x'_B, x'_N) \leq z_1(b) + \varphi(b)$ , and since this holds for all such  $(x'_B, x'_N)$ ,

$$z_2(b) \leq z_1(b) + \varphi(b). \quad (7)$$

Now let us consider the  $y$  that solves (4). For this  $y$

$$\sum_{i=1}^{i=n} c^*_{i+m} y_i = \varphi(b).$$

We extend this to a solution to (5) by choosing  $x_B = B^{-1}(b - Ny)$ . Since  $y$  solves (4),  $(b - Ny) \in \mathcal{E}_B$ , so  $x_B$  will be integral. If  $x_B$  is also nonnegative,  $(x_B, y)$  solves (1) and, in fact, is the optimal integer solution since its cost, by (6), is  $z_1(b) + \varphi(b)$ , which, by (7), establishes optimality.

To establish conditions for the nonnegativity of  $x_B$  we need the following lemma.

LEMMA. *There is an optimal solution to (4) with*

$$\sum_{i=1}^{i=n} y_i \leq D - 1.$$

*Proof:* If the  $y_i$  are the components of that optimal solution having  $\sum y_i$  minimal, form any sequence of the  $a_{i+m}$  in which each  $a_{i+m}$  appears exactly  $y_i$  times. Then form the partial sum  $S_p$  of the first  $p$  elements of the sequence. We include  $S_\delta = \delta$ . If the sequence has more than  $D - 1$  elements, there are more than  $D$   $S_p$ ; so there must be a  $p$  and  $p'$ ,  $p < p'$  for which  $S_p = S_{p'}$ . The elements, between  $p$  and  $p'$  in the sequence total  $\delta$  and can be deleted. The remaining elements form a new solution which contradicts either optimality or minimal total  $\sum y_i$ .

A related argument can be used to prove the following which, though not used in the proofs of Theorems 1, 2, or 3, bears on the multiplicity of solutions to integer programs.

**THEOREM 4.** If  $\prod_{i=1}^{i=n} (y_i + 1) > D - 1$ , then the solution to (4) is not unique.

Returning to the proof, it follows from the lemma that  $\|Ny\| \leq (D-1)l$ . Hence, if  $b \in K^s(l(D-1))$ ,  $b - Ny$  will be in  $K^s$  and so  $x_B$  will be nonnegative. This establishes Theorem 1.

Theorem 2 now follows at once from the lemma.

To establish Theorem 3, we now turn to the actual computation of  $y$ . Guided by dynamic programming,<sup>6</sup> we define  $\varphi_s(\bar{p})$ ,  $\bar{p} \in M(I)/M(B)$  as the solution to

$$\max \sum_{i=1}^{i=n} c^*_{i+m} y_i$$

$$\sum_{i=1}^{i=n} \bar{a}_{i+m} y_i = \bar{p}.$$

We have recursively

$$\varphi_s(\bar{p}) = \max \{ \varphi_s(\bar{p} - \bar{a}_{s+m}) + c^*_{s+m}, \varphi_{s-1}(\bar{p}) \}. \quad (8)$$

Let us assume that  $\varphi_{s-1}(\bar{p})$  is known for all  $\bar{p} \in M(I)/M(B)$ . Then we can compute recursively, starting with  $\varphi_s(\bar{o}) = 0$ ,

$$\varphi_s(\bar{a}_{s+m}) = \max \{ \varphi_s(\bar{o}) + c^*_{s+m}, \varphi_{s-1}(\bar{a}_{s+m}) \}$$

$$\varphi_s(r\bar{a}_{s+m}) = \max \{ \varphi_s(r\bar{a}_{s+m} - \bar{a}_{s+m}) + c^*_{s+m}, \varphi_{s-1}(r\bar{a}_{s+m}) \}, r = 1, 2, \dots, D-1.$$

If  $\bar{a}_s$  is of order  $D$ , we will obtain all values  $\varphi_s(\bar{p})$ . If  $\bar{a}_s$  is of some order  $d$  which divides  $D$ , then  $d\bar{a}_s = \bar{o}$ , and after  $d$  steps we return to  $\bar{o}$ . One then chooses some  $\bar{p}$  not yet reached in the calculation (the standard form of the group is needed here), and setting  $\varphi'_s(\bar{p}) = \varphi_{s-1}(\bar{p})$  computes  $\varphi'_s(\bar{p} + r\bar{a}_{s+m}) = \max \{ \varphi'_s(\bar{p} + r\bar{a}_{s+m} - \bar{a}_{s+m}) + c^*_{s+m}, \varphi_s(\bar{p} + r\bar{a}_{s+m}) \}$ ,  $r = 1, \dots, d$ . After  $d$  steps,  $d\bar{a}_{s+m} = \bar{o}$ , so we obtain a new value for  $\varphi'_s(\bar{p})$  and then continue obtaining new values for  $\varphi'_s(\bar{p} + r\bar{a}_{s+m})$  the second time around. As soon as one of these new values agrees with the old, the calculation is stopped. It is not hard to show that: (i) the calculation will stop after  $q$  steps  $d \leq q < 2d$ ; (ii) the  $\varphi'_s(\bar{p} + r\bar{a}_{s+m})$  values are the correct values  $\varphi_s(\bar{p} + r\bar{a}_{s+m})$ . This procedure is repeated for  $D/d$  starting points  $\bar{p}$  to get values  $\varphi_s(\bar{p})$  for all  $\bar{p} \in M(I)/M(B)$ .

If  $M \leq \min_i c^*_i$ , we can start with  $\varphi_0(\bar{p}) = M$  for all  $\bar{p}$ . Then repeating the calculation leads to the calculation of  $\varphi_s(\bar{p})$  for all  $\bar{p}$  in at most  $2nD$  elementary recursions each involving adding two group elements, looking up two values, adding two real numbers, and making one compare. To obtain the optimal solution with the smallest  $\sum_i y_i$ , one simply records with each  $\varphi_s(\bar{p})$ , when computed, the total  $T_s(\bar{p}) = \sum_i y_i$  of the  $y_i$  of that solution. Clearly,  $T_s(\bar{p}) = T_{s-1}(\bar{p})$ , if the second term gives the maximum in (8) and  $T_s(\bar{p}) = T_s(\bar{p} - \bar{a}_{s+m}) + 1$ , otherwise. In case of a tie in the maximum in (8), the term yielding the smaller  $T_s(\bar{p})$  value should be chosen.

The solutions  $y_i$  are obtained by tracing back the recursion in the usual manner of dynamic programming. By proper recording, backtracking can be done even if the  $\varphi_{s-1}$  values are discarded, once the  $\varphi_s$  are known. These backtracking operations are virtually identical with those used in solving the knapsack problem,<sup>7</sup> and, in fact,

work done with P. C. Gilmore on knapsack problems strongly suggested the results of this paper.

Finally, we note that the following steps, (i) solve  $P1$  obtaining the optimal  $B$ , (ii) put  $M(I)/M(B)$  into standard form and identify the  $a_i$ , (iii) solve (4) obtaining  $y$ , (iv) compute  $x_B = B^{-1}(b - Ny)$ , will yield an optimal solution  $(x_B, y)$  if  $x_B \geq 0$ . It is not necessary for  $b$  to be in  $K^D(l(D-1))$  to apply the procedure. The problems for which the procedure provides a solution are those for which those inequalities binding the solution of  $P1$  alone determine the solution to  $P2$ .

\* This work was supported in part by the Office of Naval Research under contract Nonr 3775-(00), NR 047040.

<sup>1</sup> Dantsig, George B., *Linear Programming and Extensions* (Princeton, N. J.: Princeton University Press, 1963).

<sup>2</sup> Gomory, R. E., "An algorithm for integer solutions to linear programs," *Recent Advances in Mathematical Programming*, ed. R. L. Graves and Philip Wolfe (McGraw-Hill, 1963), pp. 269-302.

<sup>3</sup> Gomory, R. E., "All-integer integer programming algorithm," *Industrial Scheduling*, ed. J. F. Muth and G. L. Thompson (Prentice-Hall, 1963), pp. 193-206.

<sup>4</sup> Zassenhaus, Hans, *The Theory of Groups* (New York: Chelsea Publishing Co., 1949).

<sup>5</sup> van der Waerden, B. L., *Modern Algebra* (New York: Frederick Ungar Publishing Co., 1950), vol. 2.

<sup>6</sup> Bellman, R., *Dynamic Programming* (Princeton, N. J.: Princeton University Press, 1957).

<sup>7</sup> Gilmore, P. C., and R. E. Gomory, "Multi-stage cutting stock problems of two and more dimensions," to appear in *Operations Res.*

# FACES OF AN INTEGER POLYHEDRON\*

By R. E. GOMORY

THOMAS J. WATSON RESEARCH CENTER, YORKTOWN HEIGHTS, NEW YORK

*Communicated by R. Courant, November 1, 1966*

In reference 5 a connection was given between the integer and noninteger solutions to the linear programming problem

$$\text{maximize } z = cx \quad (1)$$

$$Ax = b, x \geq 0.$$

In (1)  $x$  is an  $m + n$  vector,  $b$  is an integer  $m$ -vector,  $c$  an  $m + n$  vector, and  $A$  an  $m \times (m + n)$  integer matrix containing an  $m \times m$  identity matrix.  $A$  is assumed to be rearranged and partitioned into an  $m \times m$  optimal basis matrix  $B$  for the noninteger problem and a collection of nonbasic columns forming the matrix  $N$  with  $A = (B, N)$ . An alternative form of (1) that is useful here for geometric interpretation is to revert to inequalities, writing  $A$  as  $(A', I)$ . Then (1) becomes

$$\text{maximize } z = c'x' \quad (1a)$$

$$A'x' \leq b$$

where  $x'$  and  $c'$  are  $n$ -vectors.

Under suitable conditions, given in reference 5, the integer solution to (1) could be obtained from the noninteger one by solving the optimization problem

$$\max \sum_{i=1}^{i=n} c_i^* t_i$$

subject to the conditions

$$\sum_{i=1}^{i=n} g_i t_i = g_0 \quad (2)$$

where the  $t_i$  are required to be nonnegative integers. The  $c_i^*$  (which are not important here) are the relative cost coefficients associated with the columns of  $N$ , and  $g_i$  is the element of the factor module  $g = M(I)/M(B)$  corresponding to the  $i$ th column of  $N$ . Here  $M(I)$  is the module of all integer  $m$ -vectors, and  $M(B)$  the module generated over the integers by the columns of  $B$ .

The connection between integer and noninteger solutions established by (2) held only under certain conditions. One way to develop this approach into a general integer programming algorithm would be to develop from (2) new inequalities or "cutting planes" for a method similar to that of reference 6. The geometrical interpretation of the solutions to (2) suggests that this is possible and this approach is outlined here.

To see this, consider the cone  $P'$  in the space of the variables  $x'$  of (1a) formed by using only the inequalities corresponding to the nonbasic variables or equivalently,  $P'$  is the cone obtained if in (1a) the nonnegativity condition for the basic variables is dropped. Within  $P'$  is the polyhedron  $P''$  which is the convex hull of the integer points of  $P'$ .  $P''$  is an interesting object of study in itself. In addition, its faces

clearly provide the strongest inequalities or cutting planes for the general integer programming problem that can be deduced locally, i.e., without using the nonlocal information available from the nonnegativity condition on the basic variables.

Since the variables  $t$  of (2) determine a corresponding  $x$  satisfying the equations of (1) by  $t \rightarrow x = (B^{-1}(b - Nt), t)$  and hence also determine the  $x'$  of (1a), inequalities on the  $t_i$  yield inequalities on the  $x'$ , and in this sense one can talk about an inequality on the  $t_i$  being a face of  $P''$ .

Faces to  $P''$  can be characterized by the following easily proved theorem which allows their computation by linear programming.

**THEOREM 1.** *The inequality  $\sum \pi_i t_i \geq \pi_0$  is a face of  $P''$  if and only if the  $\pi_i$  are a basic feasible solution of the system of inequalities,*

$$\pi T \geq \pi_0 \quad (3)$$

made up from all vectors  $T = (t_1, \dots, t_n)$ , satisfying the equations of (2).

A number of remarks can be made about the feasibility of this computation.

*First:* Although there are an infinity of  $T$  satisfying (2) and hence an infinity of inequalities, it is easy to reduce this to a finite number by considering only the irreducible  $T$ , i.e., those  $T$  not containing as a vector  $T' = (t_1, \dots, t_n)$  for which  $\sum t_i q_i = 0$ . Or alternatively one can work only with those  $T$  that satisfy (2) for some nonnegative  $c_i$ .

*Second:* The trivial faces of  $P''$ , those that are simply faces of the original problem (1), can be discarded by choosing  $\pi_0 \neq 0$ .

*Third:* The multiplicity of rows in (3) can be dealt with by a row-generating method similar to the methods of references 1, 2, 4, and 7. Because of this, an  $n \times n$  basis matrix is the most that is required at any time. The needed row at each simplex step can be generated (at worst) by solving a problem approximately equivalent to (2), essentially a shortest path problem over the group  $G$ .

*Fourth:* There is a simple way of getting a first feasible solution to (3), and hence a face of  $P''$ , by solving a single problem like (2) with an additional side calculation that less than doubles the work. (For an estimate of the work involved in solving (2), see ref. 5.) This calculation will not be described here.

*Fifth:* Duplication, i.e., many columns of  $N$  mapping into the same group element, can easily be taken care of and simply reduces the size of the problem to be dealt with.

In addition to providing a method of computing faces of  $P''$  by linear programming, equations (3) lead to the proof of the following theorems which involve considering the tree of shortest paths over the group  $G$ .

Let  $g_1, \dots, g_n$  be the group elements corresponding to the columns of  $N$ . Choose from among the  $g_i$  a basis (we can assume it is  $g_1, \dots, g_p$ ) so that  $G = g_1 \oplus g_2 \oplus \dots \oplus g_p$ , the direct sum. The remaining  $g_i$  and the element  $g_0$  corresponding to  $b$  can then be represented as  $p$ -vectors with respect to this basis; so  $g_k = \sum_{i=1}^p \gamma_{k,i} g_i$  and the following theorem, whose proof is not given here, holds.

**THEOREM 2.** *If for some basis  $g_1, \dots, g_p$ , the representation of  $g_0$ ,  $g_0 = \sum \gamma_{0,i} g_i$ , has a component  $s$  for which*

$$\gamma_{0,s} \geq \max_{k > p} \gamma_{k,s},$$

then the  $\pi_i$  given by

$$\pi_0 = \gamma_{0,s}, \quad \pi_s = 1, \quad \pi_k = \gamma_{k,s}k > p, \quad \text{and} \quad \pi_k = 0$$

otherwise give a face of  $P^*$ .

The condition of the theorem is always met whenever, for some component  $s$ ,  $\gamma_{0,s}$  is exactly one less than the order of  $g_s$ . In particular, we have the corollary.

**COROLLARY.** *If  $g$  is the direct sum of cyclic groups of order 2, then any basis of  $g$  satisfies the conditions of Theorem 2.*

It is easily shown that all  $A$  consisting of columns with at most two nonzero entries that are restricted to be 1, or  $-1$ , yield groups  $G$  that are direct sums of cyclic groups of order 2. This connects with the work of Edmonds.<sup>2</sup>

In the next theorem we refer directly to the shortest path tree. In a graph, if any unambiguous method of breaking ties among paths is used so that there is a unique shortest path between two points, the shortest paths from one point to all the others will form a spanning tree. However, different tie-breaking methods produce different trees. If the elements of  $g$  are taken as nodes of a graph, and if the  $g_i$ ,  $i = 1, \dots, n$ , are taken as directed arcs connecting each  $g'$  to the point  $g' + g_i$ , we have a graph and hence for any choice of  $\pi_i$ ,  $i = 1, \dots, n$ , a shortest path tree.

In what follows,  $b'$  denotes a right-hand side in (1) and  $g'$  is the group element  $fb'$  under the natural mapping  $f$  which sends  $M(I)$  onto  $G$ . We can now state Theorem 3.

**THEOREM 3.** *If  $R$  is a shortest path tree for the  $\pi_i$  forming a face of  $P^*$ , and if  $g' = fb'$  is separated from  $\bar{0}$  in  $R$  by  $g_0$ , then the  $\pi_i$  are also a face for the polyhedron  $P^*$  resulting from the right-hand side  $b'$ .*

The  $\pi_0'$  corresponding to  $b'$  can be obtained by adding the tree distance from  $g_0$  to  $g'$  to the original  $\pi_0$ .

Finally we state a theorem that allows the computation of faces on a once-and-for-all basis, independent of the particular columns present in the matrix  $M$ .

Let the faces  $F_i$  be all faces of the higher-dimensional integer polyhedron  $P^*$  obtained from (2) by letting the index  $i$  in (2) range over all group elements.

$$F_i = (\pi_{0,i}, \pi_{1,i}, \pi_{2,i}, \dots, \pi_{D-1,i}). \quad (4)$$

**THEOREM 4.** *The faces  $F_i'$ , obtained by deleting from (4) all  $\pi_i$ , whose corresponding group element is not a column of  $N$ , include all faces of the original polyhedron  $P^*$ .*

\* This work was supported in part by the Office of Naval Research under contract Nonr 3775-(00), NR 047040.

<sup>1</sup> Dantsig, G. B., and P. Wolfe, "Decomposition principle for linear programs," *Operations Res.*, 8, 101-111 (1960).

<sup>2</sup> Edmonds, J., "Maximum matching and a polyhedron with 0, 1-vertices," *J. Res. Natl. Bur. Std.*, B, 69B, 125-130 (1965).

<sup>3</sup> Ford, L. R., Jr., and D. R. Fulkerson, "A suggested computation for maximal multi-commodity network flows," *Management Sci.*, 5, 97-101 (1958).

<sup>4</sup> Gilmore, P. C., and R. E. Gomory, "A linear programming approach to the cutting-stock problem," *Operations Res.*, 9, 849-859 (1961).

<sup>5</sup> Gomory, R. E., "On the relation between integer and noninteger solutions to linear programs," these PROCEEDINGS, 53, 260-265 (1965).

<sup>6</sup> Gomory, R. E., "An algorithm for integer solutions to linear programs," in *Recent Advances in Mathematical Programming*, ed. Robert L. Graves and Philip Wolfe (New York: McGraw-Hill, 1963), pp. 269-302.

<sup>7</sup> Gomory, R. E., and T. C. Hu, "Synthesis of a communication network," *SIAM J.*, 12, 348-369 (1964).

**MATHEMATICAL PROGRAMMING**

**by**

**CARLTON LEMKE**

**at the**

**American Mathematical Society Summer Seminar**

**on the**

**Mathematics of the Decision Sciences**

**Stanford University**

**July - August 1967**

ON  
COMPLEMENTARY PIVOT THEORY

by

C. E. Lemke

A Portion of the work was supported by  
The Air Force Office of Scientific Research  
under Research Grant No. AF-AFOSR-339-64.

Department of Mathematics  
Rensselaer Polytechnic Institute  
Troy, New York

RPI Math Report No. 76  
July 5, 1967

## ON COMPLEMENTARY PIVOT THEORY

by

C. E. Lemke

### INTRODUCTION

These notes constitute, in the main, an exposition of the results contained in References [ 4 ] and [ 5 ], and further, incorporate some thoughts of the author and results of others which have accrued since. An effort is made to make the presentation simple and instructive, to stress the reliance on and relation to well-known procedures of linear programming, and to emphasize the unifying and generalizing aspects of the results.

We are concerned with the existence and computation of solutions to the following system of constraints:

- (1)  $w = q + Mz;$
- (2)  $w \geq 0; z \geq 0;$
- (3)  $w'z = 0 (= w_1z_1 + w_2z_2 + \dots + w_nz_n);$

for various given square matrices of order  $n$ , and column  $q$ . Prime denotes matrix transposition. The above problem involves  $2n$  variables, restricted to be non-negative.  $(w_i, z_i)$  is a complementary pair; and  $w_i$  and  $z_i$  are complements of one another. A feasible solution for which at least one of each pair is equal to zero (that is,  $w_iz_i = 0$ , for each  $i$ ; which is equivalent to (3) ) is a desired complementary solution. We shall, following Cottle and Dantzig (Ref. [ 2 ]), refer to the problem of finding solutions to this system as the "Fundamental Problem".

The form is fairly general in the sense that many problems

may be so posed. Outstanding examples are (as shown in the appendices) the problem of solving the convex quadratic programming problem (which includes the linear programming problem), and the problem of finding a Nash equilibrium point for bimatrix games. The results of Section B encompass the construction and proof of solution for these and other problems. The computational schemes are 'complementary pivot schemes', which consist of a sequence of pivots starting from system (1), or from that system augmented by a single additional variable. This extends the use of such schemes.

#### A. PRELIMINARIES.

In this section we develop concisely the well-known essentials of pivotal schemes in a form most useful to the development in Section B, where the schemes and proofs are given.

One is basically concerned with a system of  $m$  linear equations in  $m+n$  unknowns:

$$(4) \quad w = q + Az; \quad (-q + z_1 A_1 + z_2 A_2 + \dots + z_n A_n)$$

where  $A$  has order  $m$  by  $n$ . Bars above variables denote explicit values of the variables. A set of  $m+n$  values  $(\bar{w}, \bar{z})$  is a solution if and only if it satisfies (4):  $\bar{w} = q + A\bar{z}$ . A solution is feasible if and only if it is non-negative:

$$(5) \quad w \geq 0; \text{ and } z \geq 0.$$

$L$  denotes the set of solutions;  $K$  the set of feasible solutions.  $K$  may be empty. Geometrically,  $L$  is a linear manifold of dimension  $n$  in  $(m+n)$ -space.  $K$  is a convex polyhedron (an intersection of half-spaces). Throughout, geometric descriptions are not necessary to understanding, and may be ignored.

In (4),  $L$  is described by giving the right-side (independent) variables values and computing the left-side (dependent) variable values. We are interested in alternative descriptions of  $L$  having the same form and obtained from (4).

$A_j$  denotes the  $j$ th column of  $A$ ;  $A_{i,j}$  denotes the component in row  $i$  and column  $j$  of  $A$ . Suppose that

$$(6) \quad A_{s,r} \neq 0.$$

We may then perform a pivot on the form (4) which consists of:  
 (i) solving the  $s$ th equation of (4) for the variable  $z_r$  (which involves a division by  $A_{s,r}$ ; called the pivot), and  
 (ii) replacing  $z_r$  by the result in each of the remaining  $m-1$  equations of (6).

The result of the pivot is again a form like (4), giving an equivalent description of  $L$ . The left-hand set of  $m$  variables differs from the previous left-hand set in one component; the pair  $(w_s, z_r)$  have been exchanged; and specifying this pair (called the pivot pair) completely specifies the pivot.

We may then perform a pivot on the result; and continue in a sequence of pivots. A pivotal scheme on (4) consists of a sequence of pivots together with criteria for (a) deciding the pivot pair on each pivot and (b) deciding when to terminate the pivoting.

A Sequence of pivots leads to an equivalent description of  $L$  given by the resulting form:

$$(7) \quad w^t = q^t + A^t z^t; \quad (-q^t + z_1^t A_1^t + z_2^t A_2^t + \dots + z_n^t A_n^t);$$

so that  $(w^t, z^t)$  is always some permutation of the variables  $(w, z)$ . We now accumulate the relevant definitions and facts.

A basic set is a set, call it  $w^t$ , of  $m$  of the  $n+m$  variables  $(w, z)$  such that  $w^t$  may be obtained from (4) in the form (7) (and by a sequence of pivots). The corresponding set  $z^t$  of the remaining  $n$  variables is the associated non-basic set. Given the basic set  $w^t$ , the form (7) is unique, except for the permuting the equations and/or the terms of right-most, expanded expression for  $w^t$ . (7) is called a basic form.

if one has:

$$(8) \quad A_{r,s}^t > 0$$

then a pivot on (7) defined by the pivot pair  $(w_s^t, z_r^t)$  yields a new basic form called adjacent to (7). Conversely, two basic forms are adjacent if and only if their basic sets differ by only one variable, (hence may be obtained, one from the other, by a single pivot). It follows that a basic form (7) has as many adjacent basic forms as there are non-zero components in  $A^t$ . Two basic sets are adjacent if their corresponding basic forms are adjacent, (the equivalence of basic forms by permuting equations or non-basic variables is always assumed).

Referring to (7), the unique point:  $(\bar{w}^t, \bar{z}^t) = (q^t, 0)$ , obtained from (7) by setting all non-basic variables  $z^t$  to zero is called the associated basic point.

A basic point has at least  $n$  zeros (since  $\bar{z}^t = 0$ ).  $L$  is non-degenerate if and only if any solution  $(\bar{w}, \bar{z})$  has at most  $n$  of the  $m+n$  values  $\bar{w}_i, \bar{z}_j$  equal to 0. Alternatively and equivalently,  $L$  is non-degenerate if and only if every basic point has exactly  $n$  zero values (exactly  $m$  non-zero

values). We assume throughout that  $L$  is non-degenerate. We shall however always make note of any effects of non-degeneracy. An immediate effect is the following:

A solution has exactly  $n$  zero components if and only if it is a basic solution; hence corresponding to each basic set is one and only one basic solution; and each basic solution corresponds to one and only one basic set. Hence, in each basic form (7)  $q^t$  has no zero components.

Two basic solutions are adjacent if their basic sets are adjacent.

Referring to (7), for fixed  $j$  all solutions satisfying:

$$(9) \quad w^t = q^t + z_j^t A_j^t$$

will be called a basic line; obtained from (7) by taking all non-basic variables except  $z_j^t$  equal to zero. Points on a basic line have either  $n$  or  $n-1$  zero components ( $m$  or  $m+1$  non-zero components). If in (9) some value of  $z_j^t$  makes a component of  $w^t$  zero, the corresponding solution has exactly  $n$  zero components; hence is a basic solution and in fact an adjacent solution to the basic solution  $(\bar{w}^t, \bar{z}^t) = (q^t, 0)$ , corresponding to (7). Hence, all adjacent solutions lie on basic lines. With reference to (9) for future reference let us (i) let  $e_j$  be a column whose  $j$ th component is '1'; whose other components are 0; (ii) write

(9) as:

$$(10) \quad \begin{pmatrix} w^t \\ z^t \end{pmatrix} = \begin{pmatrix} q^t \\ 0 \end{pmatrix} + \theta \begin{pmatrix} A_j^t \\ e_j \end{pmatrix}$$

and re-permute variables to their original order. (10) becomes

$$(11) \quad \begin{pmatrix} w \\ z \end{pmatrix} = \begin{pmatrix} \bar{w} \\ \bar{z} \end{pmatrix} + \theta \begin{pmatrix} \bar{v} \\ \bar{u} \end{pmatrix}$$

where  $(\bar{w}, \bar{z})$  is still the basic solution corresponding to (7). In order that this satisfy (4) for all  $\theta$ :

$$(12) \quad \bar{v} = A\bar{u};$$

and  $(\bar{v}, \bar{u})$  has at least  $n-1$  zero values.

Our next remarks refer to feasibility and the set  $K$ . A feasible basic form is a basic form whose basic point is feasible, called a basic feasible point. We indicate such a form by:

$$(13) \quad w^t = q^t + A^t z^t; \quad q^t > 0.$$

A basic feasible point is an extreme point of  $K$ . A feasible pivot is a pivot from one feasible basic form to another feasible basic form. We point out that corresponding to each non-basic variable  $z_j^t$  at most one feasible pivot from (13) is possible. Referring to (9), now restricting  $z_j^t$  to non-negative values, since  $q^t > 0$ ,  $z_j^t$  may be increased from 0 while retaining  $w^t \geq 0$ . Either for the first time, say for  $z_j^t = \bar{z}_j^t$  some component, say  $\bar{w}_s^t$  becomes 0; in which case the corresponding point is an adjacent feasible basic point, and a feasible pivot on (13), defined by the pivot pair  $(w_s^t, z_j^t)$  may be performed; or else solutions satisfying (9) are feasible for all  $z_j^t \geq 0$ ; which is true if and only if:

$$(14) \quad A_j^t \geq 0.$$

In the former case, solutions satisfying (9) for the interval  $0 \leq z_j^t \leq \bar{z}_j^t$  are feasible, and form a bounded edge of  $K$ , whose two end-points, defined by the end-values of the interval, are called adjacent extreme points. In the latter case,

all solutions satisfying (9) for  $z_j^t$  are feasible and form an unbounded edge of  $K$ . In particular, exactly  $n$  edges of  $K$  meet in an extreme point; and each extreme point has at most  $n$  adjacent extreme points -- each feasible basic form has at most  $n$  adjacent feasible basic forms.

A feasible pivotal scheme consists of a sequence of feasible pivots. In these notes we will be limited to feasible pivotal schemes such that, given a feasible form (13), a non-basic variable, say  $z_r^t$ , is first selected. The pivoting terminates if  $A_r^t \geq 0$ ; otherwise the pivot is uniquely defined by  $z_r^t$ . Geometrically, a feasible pivotal scheme defines an adjacent-extreme-point path.

Finally, we define a proper pivotal scheme as one for which no basic set appears twice. Since the number of basic sets is finite, a proper pivotal scheme always terminates.

B. Complementary Pivot Schemes and Applications.

We now consider the Fundamental Problem, (1), (2), (3). We consider three schemes which are essentially alike. Each scheme consists, except for one or two initial pivots, of feasible pivots. Schemes II and III apply directly to L. Scheme I requires augmenting L by one (pseudo) variable. All schemes are proper schemes; hence terminate. Following the description of each scheme; a class of problems for which the scheme applies is defined; and proofs are given.

L is now the set of solutions to (1); K is the set of non-negative solutions. L is assumed to be non-degenerate. A complementary solution; requiring  $w_1 z_1 = 0$  thus has at least n zero components; by non-degeneracy at most n zero components; hence exactly n; and hence is a basic feasible point. The number of complementary solutions is thus finite. Let C denote the set of complementary solutions.

Scheme I. Let  $z_0$  be a scalar variable;  $e > 0$  denote any column with positive components; and let  $L'$  denote the set of solutions to:

$$(15) \quad w = q + z_0 e + Mz = q + Az;$$

where  $A = (e, M)$ ;  $\underline{z} = \begin{pmatrix} z_0 \\ z \end{pmatrix}$ . Non-basic sets thus have  $n+1$  components.

Let  $K'$  be the set of feasible solutions of  $L'$ .

Let  $C_0$  denote the set of points of  $K'$  satisfying:

$$(16) \quad w'z = 0 \quad (\text{that is, } w_1 z_1 = 0; \quad i = 1, 2, \dots, n).$$

We shall generate a proper feasible sequence whose basic points satisfy (16).

Either  $q > 0$ , in which case the basic point of (15) is already a complementary solution for  $L$ , and no pivots are required; or else  $q$  has some negative component, which we assume to be the case.

On the first pivot,  $z_0$  is increased until for the first time;  $w = q + z_0 e \geq 0$ ; in which case some  $w_r$  becomes zero. The first pivot is defined by the pivot pair:

$$(17) \quad (w_r, z_0).$$

This leads to the basic form:

$$(18) \quad w^t = q^t + A^t z^t; \quad q^t > 0,$$

for  $t = 1$  The basic feasible point satisfies (16). Associated with (18) is a non-basic/complementary pair; in the case  $t = 1$  the pair  $(w_r, z_r)$ . As  $z_r$  the complement of  $w_r$  is increased in (18) (16) remains satisfied. If a pivot making  $z_r$  basic can be made; this pivot becomes the second pivot, and leads to the feasible form (18), for  $t = 2$ . If the pivot cannot be made, the sequence is terminated.

In general, suppose that  $t \geq 1$  pivots have led to the feasible form (18); and suppose that (16) was satisfied for all basic feasible points generated. If  $z_0$  is non-basic a complementary solution has been found; namely the final basic point, and the sequence terminates. If  $z_0$  is still basic, suppose that the variable that became non-basic on the  $t$ th pivot was one of the complementary pair  $(w_s, z_s)$ . Since condition (16) prevailed, both components of this pair are

non-basic. By counting, this is the only such pair -- the non-basic complementary pair associated with the current basic form.

The complement of the one of the pair which just became non-basic is to be increased. Either a unique  $t+1$ st pivot is thus specified; or the sequence is terminated. This completes a description of the scheme.

It remains to point out that the scheme generates a proper scheme; that no basic set occurs twice.

First consider (18) for  $t = 1$ .  $w_r$  has just become non-basic. We may assume that  $w_r$  is  $z_0^1$  (the first component of  $z^1$ ). By examining the first pivot it is easy to see that  $A_0^1$  (the first column of  $A^1$ ) has non-negative components. Let us label as  $E_0$  the unbounded edge of points of  $K'$  obtained for all non-negative values of  $w_r$ . Since  $\bar{z}_r = 0$  All points of  $E_0$  are in  $C_0$ ; i.e., satisfy (16).

Now for any  $t \geq 1$  consider (18); where  $z_0$  remains basic. Let  $(w_s, z_s)$  be the associated non-basic complementary pair. A feasible pivot from (18) to a feasible form satisfying (16) is only possible by increasing one or the other of this pair. When (18) is not a terminal basic form (either  $t = 1$  or a final  $t$ ) there are exactly two adjacent basic forms satisfying (16).

Therefore there can be no first basic set that repeats, for if a basic set repeats the immediately preceding basic set must have repeated. Hence the scheme is proper.

Observe that the scheme may be carried out for any  $M$  and  $q$ . Either the scheme terminates in a complementary solution of  $L$ , or in an unbounded edge of points of  $C_0$ .

In the latter case, let us label that unbounded edge  $E$ . Let us observe that  $E$  cannot be  $E_0$ . The only access to  $E_0$  is by way of the basic point of basic form (18) for  $t = 1$ ; and this form does not repeat. If (an extreme case) the sequence terminated with  $t = 1$ , then  $E$  must have been the set of points obtained from (18) by increasing  $z_r$ . In any case  $E_0$  and  $E$  are distinct.

In general, if the sequence terminates in  $E$ , no conclusions are possible. We now however isolate a class of matrices  $M$  such that termination in  $E$  implies that  $K$  is empty, for the given  $q$ .

Co-positive matrices are (square) matrices  $M$  such that:

$$(19) \quad z \geq 0 \quad \text{implies that} \quad z'Mz \geq 0.$$

Co-positive-plus matrices are co-positive matrices such that:

$$(20) \quad z \geq 0 \quad \text{and} \quad z'Mz = 0 \quad \text{imply that} \quad (M+M')z = 0.$$

The class of co-positive matrices include, most importantly, positive-semi-definite matrices ( $z'Mz \geq 0$ ; all  $z$ ), and non-negative matrices ( $M \geq 0$ ; that is, all components non-negative). There are also ways of compounding co-positive matrices. For example, a non-negative linear combination of co-positive matrices is co-positive.

The class of co-positive-plus matrices also includes positive-semi-definite matrices, and the class of strictly co-positive matrices ( $z \geq 0$ , and  $z \neq 0$  imply  $z'Mz > 0$ ); which includes positive matrices ( $M > 0$ ). Again, co-positive-plus matrices may be compounded to form co-positive matrices.

For example, positive linear combinations of co-positive-plus matrices are co-positive-plus. If  $M_1$  and  $M_2$  are co-positive-plus, then so is  $M_1'$ , and, since anti-symmetric matrices ( $M' = -M$ ; or, equivalently,  $z'Mz = 0$ ; all  $z$ ) are positive-semi-definite, as observed by Cottle (see Ref.[2]), the matrix  $\begin{pmatrix} M_1 & -A' \\ A & M_2 \end{pmatrix}$  is co-positive-plus, for any  $A$ .

In particular, to ascertain whether or not  $M$  is co-positive-plus it is sufficient to test its symmetric part  $\frac{1}{2}(M+M')$ ; since one is co-positive-plus if and only if the other is. Thus a study of the class could be restricted to the symmetric sub-class: co-positive symmetric matrices for which:  $z \geq 0$  and  $z'Mz = 0$  imply that  $Mz = 0$ .

Let  $C_+$  denote the class of co-positive-plus matrices. (The term 'co-positive-plus' and notation ' $C_+$ ' were suggested by Cottle.)

Theorem 1. Let  $M$  be in  $C_+$ . Then Scheme I terminates either in a complementary solution of  $L$  or in the conclusion that for the given  $q$  no feasible solution exists.

Proof: We will suppose that the Scheme terminated in the unbounded set  $E$  contained in  $C_0$ , and show that then there is no feasible solution. Let (18) designate the final basic form. We are supposing that, for some  $r$ , all points of  $L$  satisfying:

$$(21) \quad w^t = q^t + \underline{z}_r^t A_r^t \quad \text{for} \quad \underline{z}_r^t \geq 0;$$

where points satisfying (16); where  $A_r^t \geq 0$ .

By (10), (11), (12)  $E$  may be described as the set

$$(22) \quad w = \bar{w} + \theta \bar{v}; \quad z = \bar{z} + \theta \bar{u}; \quad z_0 = \bar{z}_0 + \theta \bar{u}_0,$$

for  $\theta \geq 0$ ; where  $(\bar{z}, \bar{w})$  is the basic feasible point corresponding to the final basic form.

We draw the following conclusions:

- a. Since  $E$  is assumed to exist; both Case 1:  $\bar{u} = 0$  and Case 2:  $\bar{u}_0 = 0$  cannot hold. We show that Case 1 cannot hold and that Case 2 must hold.
- b. For  $w, z, z_0$  to satisfy (15) for all  $\theta \geq 0$  requires:

$$(23) \quad \bar{v} = \bar{u}_0 e + M\bar{u}.$$

- c. For  $w, z, z_0$  to remain non-negative for all  $\theta \geq 0$  requires:

$$(24) \quad \bar{u} \geq 0; \bar{v} \geq 0; \bar{u}_0 \geq 0.$$

- d. That  $w'z = 0$  hold for all  $\theta > 0$  requires (since all terms in (22) are non-negative):

$$(25) \quad \bar{z}'\bar{w} = \bar{z}'\bar{v} = \bar{w}'\bar{u} = \bar{u}'\bar{v} = 0.$$

If  $\bar{u} = 0$ : (23), (24) imply that  $\bar{v} = \bar{u}_0 e > 0$ . By (25):  $\bar{z}'\bar{v} = 0$  implies that  $\bar{z} = 0$ . Hence  $w = z_0 e$  describes  $E$ . But this is precisely the set  $E_0$ . Hence  $\bar{u} \neq 0$ . For simplicity we normalize  $\bar{u}$  by:  $e'\bar{u} = 1$ .

We may write:

$$(26) \quad z'w = z'Mz + z'(q + z_0 e) = 0.$$

- e. Now suppose that  $M$  is co-positive. By (26):

$$(27) \quad f(\theta) = -z'(q + z_0 e) - z'Mz \geq 0;$$

where  $f(\theta)$  is quadratic in  $\theta$ . In order for (27) to hold for all  $\theta \geq 0$ , requires that the coefficient of  $\theta^2$  be non-negative:

$$(28) \quad -\bar{u}_0(\bar{u}'e) = -\bar{u}_0 \geq 0;$$

which is possible only if  $\bar{u}_0 = 0$ .

Thus  $f(0)$  is linear in  $0$ , and for (27) to hold for all  $0 \geq 0$  requires that the coefficient of  $0$  be non-negative:

$$(29) \quad -\bar{u}'(q + \bar{z}_0)e \geq 0; \text{ or: } \bar{u}'q \leq -\bar{z}_0 < 0;$$

the latter inequality since we are assuming that  $z_0$  was basic at the final basic form.

Summarizing, if  $M$  is co-positive, the assumption of  $E$  implies a column  $\bar{u}$  satisfying:

$$(30) \quad M\bar{u} \geq 0; \quad \bar{u} \geq 0; \quad \bar{u}'q < 0; \text{ and } \bar{u}'M\bar{u} = c.$$

f. Finally, assume that  $M$  is co-positive-plus. Then  $M'\bar{u} - M\bar{u} \leq 0$ . Then for any  $z \geq 0$  and  $w = q + Mz$  we have:

$$(31) \quad \bar{u}'w = \bar{u}'q + z'(M'\bar{u}) < 0.$$

Hence  $w$  is not non-negative; that is,  $K$  is empty.

This completes the proof.

Scheme II. For this scheme, we shall assume that  $M$  has a positive column. We work directly with the Fundamental Problem. We may assume that the first column of  $M$  is positive. Again if  $q > 0$ , the initial basic solution is a complementary solution. We assume that  $q$  has a negative component. Then increasing  $z_1$  defines a unique first pivot defined by the pivot pair  $(w_r, z_1)$ ; for some  $r$ ; leading to the basic form:

$$(32) \quad w^t = q^t + M^t z^t; \quad q^t > 0;$$

for  $t = 1$ . This has a basic feasible solution satisfying:

$$(33) \quad w'z = w_1 z_1 \quad (\text{i.e.; } w_i z_i = 0 \text{ for } i \neq 1)$$

Now let  $C_1$  be the set of points of  $K$  satisfying (33). (Dantzig has called such points 'almost complementary' points).

Entirely analogous to Scheme I, Scheme II pivots so as to retain (33). This defines a proper sequence of pivots. Since points obtained from (1) for  $z_1$  large; other  $z_j$  kept at zero; are points of  $C_1$ , increasing  $w_r$  in (32) (with  $t = 1$ ) defines the initial unbounded edge  $E_0$  consisting of points of  $C_1$ . Again, associated with each whose basic solution is not complementary basic form (32)/for  $t \geq 1$ , is a non-basic complementary pair, say  $(w_s, z_s)$ ; one of which became non-basic on the  $t$ th pivot, and the other of which is to be increased to define a  $(t+1)$ st pivot if possible. If the pivot is not possible, another unbounded edge  $E$  of  $C_1$  is identified, and the sequence of pivots terminates.

Hence, again, the sequence terminates either in a complementary solution, or in the unbounded edge  $E$ ; distinct from  $E_0$ .

The case  $M > 0$  furnishes an example of a class for which both Schemes I and II terminate in a complementary solution for any  $q$ , since  $M$  is in  $C_+$ , and since there is no question of feasibility. The proof that Scheme II works is extremely simple:

Theorem 2. If  $M > 0$ , Scheme II terminates in a complementary solution for any  $q$ .

Proof: We point out that the set  $C_1$  contains only the unbounded edge  $E_0$  and no other. Hence that the pivots cannot terminate in an unbounded edge. Indeed, if

$$(34) \quad w = \bar{w} + \theta \bar{v}; \quad z = \bar{z} + \theta \bar{u}; \quad \theta \geq 0$$

are points satisfying (33), and are feasible, so that:

$\bar{v} = M\bar{u}$ ;  $\bar{u}, \bar{v} \geq 0$ ; and  $\bar{u} \neq 0$ ; then  $\bar{v} > 0$ , since  $M > 0$ .

For all  $i \neq 1$ , by (33):  $0 = w_i(\bar{z}_i + \theta \bar{u}_i)$ ; and since

$w_i = \bar{w}_i + \theta \bar{v}_i > 0$  for  $\theta > 0$ , we have:  $\bar{z}_i = \bar{u}_i = 0$ .

Hence, on the assumed unbounded edge (34) only  $z_1$  is different from zero; hence the edge is  $E_0$ . Hence, Scheme II terminates in a complementary solution.

A slightly different situation arises in the case of a bi-matrix game example, defined as an  $M$  of the form:

$$(35) \quad M = \begin{pmatrix} 0 & A \\ B & 0 \end{pmatrix}; \quad \text{where } A > 0; \quad B > 0.$$

In this case,  $M$  is co-positive, but is not in  $C_+$ , since for  $z = \begin{pmatrix} z_1 \\ 0 \end{pmatrix}$ ; where  $z_1 \neq 0$ ; it is true that  $z'Mz = 0$ , but  $(M+M')z = \begin{pmatrix} 0 \\ (B+A')z_1 \end{pmatrix} \neq 0$ .

Indeed, for  $q$  of the form:  $q = \begin{pmatrix} -e_1 \\ e_2 \end{pmatrix}$ ; where the  $e_i$  are positive columns, (1) may be written as two formats:

$$Az_2 - e_1 = w_1; \quad Bz_1 + e_2 = w_2.$$

Then  $L$  is always feasible; but requires that  $w_2 > 0$  and  $z_2 \neq 0$ , so that  $w_2'z_2 = 0$  cannot be satisfied. Hence Scheme I cannot terminate in a complementary solution in this case.

However, the case where  $q = -e$  ( $e > 0$ ), of application to bimatrix games, is an example of a pair  $M, q$  for which a complementary solution always exists. A first constructive proof of this was given in Ref.[4], where the type of pivotal scheme considered in this section was first devised. The scheme is almost the same as Scheme II. It requires, as a major difference, two initial pivots to achieve feasibility.

We shall first describe the Scheme, and then prove the associated theorem.

Scheme III. In the case in point,  $L$  may be expressed in the disjoint forms:

$$(36) \quad \begin{aligned} \text{I:} \quad & u = -e + Ay; \quad u \geq 0; \quad y \geq 0 \\ \text{II:} \quad & v = -e + Bx; \quad v \geq 0; \quad x \geq 0 \end{aligned}$$

The complementary solution condition is:

$$(37) \quad u'x = 0; \quad \text{and} \quad v'y = 0;$$

thus;  $(u_i, x_i)$  are complementary pairs; and  $(v_j, y_j)$  are. Although the two positive  $e$ 's have different order, no confusion need arise. Scheme III consists of the following sequence:

first pivot: In I, increase  $y_1$  to obtain I-feasibility; that is, pivot on the pivot pair  $(u_r, y_1)$ ;  $r$  automatically defined.

second pivot: having determined  $r$  on the first pivot; in II increase  $x_r$  (complement of  $u_r$ ) to obtain initial II-feasibility. Let  $(v_s, x_r)$  be the pivot pair.

Then, if  $s = 1$  the resulting basic points satisfy (37), and the pivoting terminates. If not, they continue.

Let  $C_1$  denote the set of points of  $K$  satisfying:

$$(38) \quad u'x + v'y = v_1 y_1.$$

In any case, the point  $\bar{z} = \left( \frac{\bar{x}}{\bar{y}} \right) /$  satisfies (38) where  $\bar{y}$  and  $\bar{x}$  are the basic points of I and II respectively obtained by the first two pivots,

The scheme consists of alternating pivots on I and II; always increasing the complement of the variable which,

on the immediately-previous pivot became non-basic. By this scheme pivoting is automatic, and all basic points  $z$  encountered are in  $C_1$ .

**Theorem III** Scheme III terminates in a complementary solution.

**Proof:** As in Theorem 2, we may show that an initial unbounded edge  $E_0$  contained in  $C_1$  is the only unbounded edge contained in  $C_1$ .

Let  $K_I$  denote the set of I-feasible points, and similarly  $K_{II}$  for II. Associated with each basic point  $\bar{w} = \begin{pmatrix} \bar{u} \\ \bar{v} \end{pmatrix}$ ;  $z = \begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix}$  of  $K$  are the basic points  $(\bar{u}, \bar{y})$  of  $K_I$  and  $(\bar{v}, \bar{x})$  of  $K_{II}$ , and during a pivot only one of these latter points is changed. In particular, points on an unbounded edge of  $K$  must be either of the form:  $(w, z) = \left( \begin{pmatrix} u \\ \bar{v} \end{pmatrix}, \begin{pmatrix} \bar{x} \\ y \end{pmatrix} \right)$ ; where  $(\bar{v}, \bar{x})$  is a basic point of  $K_{II}$ , and  $(u, y)$  lies on an unbounded edge of  $K_I$ ; or else of the form  $(w, z) = \left( \begin{pmatrix} \bar{u} \\ v \end{pmatrix}, \begin{pmatrix} x \\ \bar{y} \end{pmatrix} \right)$ ; where  $(\bar{u}, \bar{y})$  is a basic point of  $K_I$ ; and  $(v, x)$  lies on an unbounded edge of  $K_{II}$ . (This is true, since points on an unbounded edge of  $K$  have  $n$  or  $n-1$  zero components.)

The unbounded edge  $E_0$  is identified as the set of points  $z = \begin{pmatrix} x \\ \bar{y} \end{pmatrix}$ ; obtained from the first basic point by allowing  $x_r$  (complement of  $u_r$ ) to go to infinity.

Now, let  $E$  be any unbounded edge assumed to be in  $C_1$ . We show (Case i) that  $E$  cannot involve an unbounded edge of  $K_I$  and (Case ii) if  $E$  involves an unbounded edge of  $K_{II}$ , then  $E$  is  $E_0$ .

For Case i,  $E$  is a set of points of the form:

$$(39) \quad y = \bar{y} + \theta \hat{y}; \quad u = \bar{u} + \theta \hat{u}; \quad x = \bar{x}; \quad v = \bar{v}; \quad \text{and} \quad y \neq 0.$$

For (38) to hold for all  $\theta \geq 0$  requires that:

$$(40) \quad \hat{u} = A\hat{y}; \quad \hat{u} \geq 0; \quad \hat{y} \geq 0; \quad \text{and} \quad \hat{u}'\bar{x} = 0.$$

But  $\hat{y} \neq 0$  and  $A > 0$  imply that  $\hat{u} > 0$ . Then  $\hat{u}'\bar{x} = 0$  implies that  $\bar{x} = 0$ ; which is impossible, since  $\bar{x}$  is II-feasible.

In Case ii,  $E$  is the set of points of the form:

$$(41) \quad y = \bar{y}; \quad u = \bar{u}; \quad x = \bar{x} + \theta \hat{x}; \quad v = \bar{v} + \theta \hat{v}; \quad \hat{x} \neq 0.$$

And that (38) hold for all  $\theta \geq 0$  requires:

$$(42) \quad \hat{v} = B\hat{x}; \quad \hat{v} \geq 0; \quad \hat{x} \geq 0 \quad \text{and} \quad \hat{v}'\bar{y} = \hat{v}_1\bar{y}_1; \quad \bar{u}'\hat{x} = 0.$$

But  $\hat{x} \neq 0$  and  $B > 0$  imply that  $\hat{v} = B\hat{x} > 0$ . Then  $\hat{v}_1\bar{y}_1 = 0$  for  $i \neq 1$  implies that  $\bar{y}_i = 0$ ; that is  $\bar{y}_i = 0$  for  $i \neq 1$ . That is,  $\bar{y}$  has only one positive component; namely  $\bar{y}_1$ . Also, since  $(\bar{u}, \bar{y})$  is a basic point of II, and  $\bar{y}$  has only the positive component  $\bar{y}_1$ , all components of  $\bar{u}$  except one must be positive. But, by the first pivot, this component must be  $u_r$ :  $\bar{u}_r = 0$ . Then the condition  $\bar{u}'\hat{x} = 0$  requires that  $\bar{x}_i = \hat{x}_i = 0$ ; for  $i \neq r$ ; that is, only  $x_r$  is positive in (41). This identifies the unbounded edge  $E$  as  $E_0$ , and completes the proof.

We shall terminate this section with an observation and some remarks.

Referring to the Fundamental Problem (1), (2), (3), consider the set  $C_1$  of all points of  $K$  satisfying:

$$(43) \quad w'z = w_1z_1 \quad (\text{that is, } w_jz_j = 0 \text{ for } j \neq 1).$$

$C_1$  may be empty of course. Assuming that  $L$  is non-degenerate, if  $(\bar{w}, \bar{z})$  is in  $C_1$  (43) requires that the point have at

least  $n-1$  zeros; hence by non-degeneracy either  $n$  or  $n-1$  zeros. In the former case the point is basic feasible; in the latter case it lies on an edge of  $K$ . In the former case, the extreme point is either a complementary solution or is not. If it is a complementary solution then, in the corresponding basic form just one of the pair  $(w_i, z_i)$  is non-basic, and only that non-basic variable may be increased to remain in  $C_i$ . If the extreme point is not a complementary solution, then both members of the complementary pair  $(w_i, z_i)$  are basic in the corresponding basic form; and therefore, to satisfy (43) the other  $n-2$  basic variables must have their complements non-basic. Therefore, the other two non-basic variables must be a non-basic pair  $(w_s, z_s)$  say; and making just one of these positive is the only way to move from the extreme-point; that is, for a non-complementary basic point of  $C_i$  exactly two edges contained in  $C_i$  meet in the point; and moving along one of these edges leads either to another extreme point in  $C_i$  or along an unbounded edge of points of  $C_i$ .

We therefore conclude that  $C_i$  is the union of a finite number of disjoint adjacent-extreme-point paths (which are 'proper'; that is, no extreme point on the path has more than two edges joining it; and has one edge if and only if it is a complementary point). And  $C$  (the set of complementary points) is precisely the set of end-points of the paths.

It is also clear that if  $i \neq i'$ , a point in both  $C_i$  and  $C_{i'}$  must be a complementary solution. Hence  $C$  is the intersect of the sets  $C_i$ ;  $i = 1, 2, \dots, n$ .

We have used these 'almost complementary' sets in the preceding schemes.

An enumeration of the possible kinds of paths in a set  $C_i$  brings out the fact that the path may contain (i) no complementary solution (in the case where the path has no end-points; that is, either a closed path or one which contains two unbounded edges); (ii) two complementary solutions (when the path has two end-points; i.e., is not closed and has no unbounded edges); and (iii) one complementary solution (one end-point and one unbounded edge). In the previous schemes favorable paths were of the form (iii).

Regarding such schemes, postulating for example that, for a given  $M$  and  $q$ ,  $C_i$  for some  $i$  has only one unbounded edge of  $K$  ensures the existence of a complementary solution (for example, the case  $M > 0$ , and Scheme II).

Also, for the case of an  $M$  and  $q$  for which there is a unique complementary solution, since precisely  $n$  edges of  $K$  meet in this point, exactly one of the  $n$  edges is in  $C_i$  for each  $i$ , and the paths of  $C_i$  containing that edge must have (at the other end of the path) an unbounded edge. Therefore one may conclude that such a  $K$  must contain at least  $n$  unbounded edges. An example is given by:

**Theorem 4:** If  $M$  is positive-semi-definite, and  $L$  is non-degenerate and  $K$  is non-empty, then there is exactly one complementary solution.

**Proof:** If  $(\bar{w}, \bar{z})$  and  $(\bar{w}^*, \bar{z}^*)$  are complementary solutions, so that  $\bar{w}'\bar{z} = \bar{w}^*\bar{z}^* = 0$ ; then  $\bar{w}^* - \bar{w} = M(\bar{z}^* - \bar{z})$  and:

$$-(\bar{w}^* \bar{z} + \bar{w}' \bar{z}^*) = (\bar{z}^* - \bar{z})' (\bar{w}^* - \bar{w}) = (\bar{z}^* - \bar{z})' M (\bar{z}^* - \bar{z}) \geq 0;$$

which implies that  $\bar{w}^* \bar{z} = \bar{w}' \bar{z}^* = 0$ ; which implies that

$(\bar{w}^* + \bar{w})' (\bar{z}^* + \bar{z}) = 0$ ; hence  $(\bar{w}^* + \bar{w}, \bar{z}^* + \bar{z})$  have at least  $n$  zeros; hence at most  $n$  zeros, since  $(\bar{w}, \bar{z})$  is basic.

We must conclude that  $(\bar{w}, \bar{z})$  and  $(\bar{w}^*, \bar{z}^*)$  have the same positive components; that is, the same basic set. Hence, since a basic set has just one basic point associated with it,  $(\bar{w}, \bar{z}) = (\bar{w}^*, \bar{z}^*)$ .

### C. CONCLUDING REMARKS.

The use of 'Gauss-Jordan pivots' has a long history; certainly in the solution of systems of linear equations, and more recently in Mathematical Programming, starting with the simplex method of G. B. Dantzig. The format (1), (2) is essentially the format for the simplex method. It has been exploited extensively by A. W. Tucker (see Ref. [10] ), and others since.

The Fundamental Problem probably assumed importance as a form for solving the convex quadratic programming problem. Essentially all methods proposed to date for this problem may be approached from this form. The extension of the use of the form of the problem and the applicability of similar schemes to bimatrix games (a 'non-convex' example) lent more importance to the Fundamental Problem. Examples from sources other than Programming and Game Theory are noted by Cottle and Dantzig (Ref. [2] ).

By way of extending the class of problems which have a complementary solution, besides the classes mentioned in these notes the classes of 'adequate' matrices considered by Singleton (Ref. [8] ), and matrices with 'positive principal minors' (p.p.m.) considered by Cottle and Dantzig, and by A. W. Tucker (see Ref. [10] ) are worth mentioning.

It may be noted that Theorems 1-3 in these notes rely on the particular schemes I-III; which, although reasonably 'natural' do not preclude the discovery and development of equally valuable pivotal schemes, applicable to these or other classes of problems. For example, extending the class for which Scheme I is appropriate, Johnson (Ref. [6] ) has

observed that matrices which are 'principal pivotal transforms' (see Ref.[9] ) of co-positive-plus matrices could also be handled by Scheme I. As another example, the algorithm of Cottle and Dantzig (see Ref.[1]), predating the development of the Scheme I, exploiting principal pivot transforms, may be used on the class of p.p.m. matrices. Parsons (Ref. [6] ) has developed a principal pivoting scheme for the convex quadratic programming problem, and has also pointed out that, for this problem, (more generally for positive-semi-definite M) Scheme I may be described in terms of principal pivots. These examples are certainly not exhaustive.

Scarf (see References), in a series of Cowles Commission papers, has extended the use of the 'almost complementary' paths with potentially valuable results.

## APPENDICES

### Appendix 1. The Convex Quadratic Programming Problem.

A usual statement of this problem is:

$$(44) \quad \text{Minimize } c'x + \frac{1}{2}x'Qx; \text{ where } Ax \geq b; \quad x \geq 0.$$

A general form of the linear programming problem is obtained by taking  $Q = 0$ . The functional is convex; which is the case if and only if  $Q$  is a positive-semi-definite matrix (which may be considered symmetric). The general 'Kuhn-Tucker' conditions (see Ref.[3]) in order that an  $x$  solve (44) become necessary and sufficient conditions. These conditions, when written concisely, become: find  $x, y, u$ , and  $v$  satisfying:

$$(45) \quad \begin{pmatrix} 0, & A \\ -A', & Q \end{pmatrix} \begin{pmatrix} y \\ x \end{pmatrix} = \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} b \\ -c \end{pmatrix}; \quad \begin{pmatrix} y \\ x \end{pmatrix}, \begin{pmatrix} u \\ v \end{pmatrix} \geq 0; \quad \begin{pmatrix} y \\ x \end{pmatrix}' \begin{pmatrix} u \\ v \end{pmatrix} = 0$$

which is an example of the Fundamental Problem. In this example:  $M = \begin{pmatrix} 0, & A \\ -A', & Q \end{pmatrix}$ ; a positive-semi-definite matrix;

hence co-positive-plus. Therefore, Theorem 1 gives a constructive proof of the existence of a solution to (45), (and of the duality theorem of linear programming).

### Appendix 2. Equilibrium points of Bimatrix Games.

Mixed strategies are columns  $x$  and  $y$  satisfying:  $e'x = 1; x \geq 0$  and  $e'y = 1; y \geq 0$ ; where ' $e$ ' denotes a column of appropriate order with components all 1. A Nash equilibrium point, for given  $r$  by  $s$  matrices  $A$  and  $B$  is defined (see, for example, Ref[4]) as a pair  $(\bar{x}, \bar{y})$  of mixed strategies such that for all mixed strategies

$(x, y):$

$$(46) \quad \begin{aligned} \bar{x}'\bar{A}\bar{y} &\leq x'\bar{A}\bar{y} \\ \bar{x}'\bar{B}\bar{y} &\leq \bar{x}'B y \end{aligned}$$

$x'\bar{A}\bar{y}$  is here interpreted as the 'long-range loss per play to Player I' if he plays each play according to probabilities  $x$  and Player II plays according to probabilities  $y$ .

Let  $e_i$  be a column with  $j$ th component 1; and other components zero. The  $e_j$  are mixed strategies called pure strategies. It is readily seen that (46) holds if and only if it holds for all pure strategies:  $(x, y) = (e_i, e_j)$ . In vector form the equivalent condition becomes:

$$(47) \quad \begin{aligned} (\bar{x}'\bar{A}\bar{y})e &\leq \bar{A}\bar{y} \\ (\bar{x}'\bar{B}\bar{y})e &\leq B'\bar{x} \end{aligned}$$

where  $e$  has all 1's. If  $E = ee'$  is a matrix of all 1's,  $1 = (x'e)(e'y) = x'Ey$ . Therefore, adding any multiple of  $E$  to  $A$  and  $B$  in (46) does not change the set of equilibrium points. We may therefore assume that  $A > 0$ ;  $B > 0$ . Hence  $x'\bar{A}\bar{y}$  and  $x'B y$  are positive for mixed strategies  $(x, y)$ . Consider constraints of the form:

$$(48) \quad \begin{aligned} \theta_1 e &\leq \bar{A}\bar{y} & \bar{x}'(\bar{A}\bar{y} - \theta_1 e) &= 0 \\ \theta_2 e &\leq B'\bar{x} & \bar{y}'(B'\bar{x} - \theta_2 e) &= 0. \end{aligned} \quad \text{and}$$

Since for mixed strategies  $(\bar{x}, \bar{y})$ , the 'complementary' constraints on the right are simply:  $\theta_1 = \bar{x}'\bar{A}\bar{y}$ ;  $\theta_2 = \bar{x}'B\bar{y}$ , the above system of constraints is equivalent to (47).

Next, defining variables  $x = \bar{x}/\theta_2$ ;  $y = \bar{y}/\theta_1$ ; and introducing slacks, (48) becomes:

$$(49) \quad \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} -e \\ -e \end{pmatrix} + \begin{pmatrix} 0 & A \\ B' & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}; \quad \begin{pmatrix} u \\ v \end{pmatrix}, \begin{pmatrix} x \\ y \end{pmatrix} \geq 0; \quad \begin{pmatrix} x \\ y \end{pmatrix}' \begin{pmatrix} u \\ v \end{pmatrix} = 0,$$

a problem in the 'fundamental form' where, in this case:

$M = \begin{pmatrix} 0 & A \\ B' & 0 \end{pmatrix}$ . Conversely, since for  $x$  and  $y$  to yield feasible  $u$  and  $v$  requires that  $x \neq 0$ ;  $y \neq 0$ , if  $(x, y)$  is a solution to (49),  $\bar{x} = x/x'e$ ;  $\bar{y} = y/e'y$  yields a solution to (48), so that (48) and (49) are equivalent.

Appendix 3. In Section A we have remarked that a 'pivotal scheme' is completely determined when the criteria for selecting the pivot pair and for terminating are given. The following is an example of pedagogical value in developing the basic elements of linear algebra, not only in isolating all computations as 'pivot schemes', but basing the usual theorems (relating to independence, rank, etc.) thereon. (In this regard, see also the remarks of A. W. Tucker, Ref. [10]). There is nothing new in the computations; nor indeed in the idea, but perhaps in the presentation.

Consider finding all solutions to a linear system:

$$(50) \quad 0 = q + Az.$$

Ultimately, this means 'solve (50) for some of the variables in terms of the remaining ones'. Introduce 'pseudo' variables  $w$  and consider pivots on:

$$(51) \quad w = q + Az.$$

Pivot according to the following scheme: having obtained a basic form, examine the current non-basic set. If it contains

an original  $z_j$ , say  $z_j = z_s^t$ , examine the column coefficient  $A_s^t$ . If  $A_{r,s}^t \neq 0$  for some  $r$  such that  $w_r^t$  is a pseudo variable, perform the pivot determined by the pivot pair  $(w_r^t, z_s^t)$ . When no such  $z_r^t$  or no such  $A_{r,s}^t$  can be found terminate. The iterations terminate in  $R$  pivots, where  $R = \text{Rank } A$ . The final basic form gives all solutions to (50) by setting all pseudo variables equal to 0.

#### REFERENCES

- (1) Dantzig, G. B., and Cottle, R. W., Positive Semi-definite Programming, revised in Nonlinear Programming (J. Abadie, Ed.) North-Holland, Amsterdam, 1967, pp 55-73.
- (2) Cottle, R. W. and Dantzig, G. B., Complementary Pivot Theory of Mathematical Programming, Tech Rep No. 67-2 OR House, Stanford Univ., Calif. April 1967.
- (3) Kuhn, H. W., and Tucker, A. W., Non-linear Programming, Proc. 2nd Berkeley Symp on Math. Stat. & Probability, Univ. of Calif. Press (1950).
- (4) Lemke, C. E., and Howson, J. T., Equilibrium Points of Bimatrix Games; SIAM Journal, Vol. 12, July, 1964.
- (5) Lemke, C. E., Bimatrix Equilibrium Points and Mathematical Programming, Management Sciences, Vol 11, 7, May '65.
- (6) Parsons, T. D., A Combinatorial Approach to Convex Quadratic Programming; Doctoral dissertation, Dept. of Mathematics, Princeton, University, May 1966.
- (7) Scarf, H. E., Cowles Foundation Discussion Papers.  
The Core of an n-Person Game, (revised), CFDP 182.  
An Algorithm for a Class of Non-convex Programming Problems, CFDP 211, 14 Jul 66.  
The Approximation of Fixed Points of a Continuous Mapping. CFDP 216, 23 Nov 66.

- (8) Singleton, A. W., A Problem in Linear Inequalities;  
Proc. London Math. Soc., 16, 1966, pp 519-536.
- (9) Tucker, A. W., Principal Pivotal Transforms of Square  
Matrices, SIAM Review 5, 1963, p.305.
- (10) Tucker, A. W., Combinatorial Theory Underlying Linear  
Programs, in Recent Advances in Mathematical Pro-  
gramming., (R. L. Graves, and P. Wolfe, Eds.),  
McGraw-Hill, 1963.

**OPTIMAL INVENTORY CONTROL**

by

**A. F. VEINOTT, JR.**

at the

**American Mathematical Society Summer Seminar**

on the

**Mathematics of the Decision Sciences**

**Stanford University**

**July - August 1967**

Reprinted with permission from

J. SIAM APPL. MATH.  
Vol. 14, No. 5, September, 1966  
Printed in U.S.A.

Copyright 1966 by  
Society for Industrial  
and Applied Mathematics.  
All rights reserved.

## ON THE OPTIMALITY OF $(s, S)$ INVENTORY POLICIES: NEW CONDITIONS AND A NEW PROOF\*

ARTHUR F. VEINOTT, JR.†

**Abstract.** Scarf [6] has shown that the  $(s, S)$  policy is optimal for a class of discrete review dynamic nonstationary inventory models. In this paper a new proof of this result is found under new conditions which do not imply and are not implied by Scarf's hypotheses. We replace Scarf's hypothesis that the one period expected costs are convex by the weaker assumption that the negatives of these expected costs are unimodal. On the other hand we impose the additional assumption not made by Scarf that the absolute minima of the one period expected costs are (nearly) rising over time. For the infinite period stationary model, this last hypothesis is automatically satisfied. Thus in this case our hypotheses are weaker than Scarf's. The bounds on the optimal parameter values given by Veinott and Wagner [12] are established for the present case. The bounds in a period are easily computed, and depend only upon the expected costs for that period. Moreover, simple conditions are given which ensure that the optimal parameter values in a given period equal their lower bounds. When there is no fixed charge for ordering, this reduces to earlier results of Karlin [5] and Veinott [9], [10], [11] for the nonstationary case. The above result is exploited to extend the planning horizon theorem of Veinott [9] to the case where there is a fixed charge for ordering.

**1. Model formulation.** We consider a single product dynamic inventory model in which the demands  $D_1, D_2, \dots$ , for a single product in periods  $1, 2, \dots$ , are independent random variables with distributions  $\Phi_1, \Phi_2, \dots$ . Assume  $\{\eta_i\}$  are given constants such that<sup>1</sup>  $D_i \geq \eta_i$  for all  $i$ . At the beginning of each period the system is reviewed. An order may be placed for any nonnegative quantity of stock. An order placed at the beginning of period  $i$  is delivered at the beginning of period  $i + \lambda$ , where  $\lambda$  is a known non-negative integer.

Let  $x_i$  denote the stock on hand and on order prior to placing any order in period  $i$ . Let  $y_i$  denote the stock on hand and on order after ordering in period  $i$ . It is possible for  $x_i$  and  $y_i$  to be negative indicating the existence of a backlog. We assume that the amount of stock on hand and on order at the end of period  $i$  is a specified Borel function  $v_i(y_i, D_i)$  of  $y_i$  and  $D_i$ .

\* Received by the editors July 26, 1965, and in revised form December 27, 1965.

† Program in Operations Research, Stanford University, Stanford, California. This research was supported by the Office of Naval Research under Contract Nonr-225(77) and by a grant from the Western Management Science Institute.

<sup>1</sup> Actually the main results of the paper given in §2 also hold in the more general case where  $D_i$  is a random vector. All that is required is to let  $\mathcal{D}_i$  be the Borel set of possible values of  $D_i$ , and replace the interval  $[\eta_i, \infty)$  of possible values of  $D_i$  everywhere by  $\mathcal{D}_i$ . This more general formulation allows for consideration of several classes of demands, random deterioration rates, random departures of backlogged demand, and random prices, for example, by suitable interpretation of the components of  $D_i$ .

Thus  $x_{i+1} = v_i(y_i, D_i)$ . If  $\lambda > 0$  we assume that all unsatisfied demand is backlogged so  $v_i(y_i, D_i) = y_i - D_i$ .

When  $\lambda = 0$  our formulation provides for the possibilities of deterioration of stock in storage (perishable goods) and partial backlogging of unsatisfied demand [11, p. 766]. For example suppose that whenever  $y_i < D_i$ , then a fraction  $b$  ( $0 \leq b \leq 1$ ) of the unsatisfied demand is backlogged and the remainder leaves immediately. If instead  $y_i > D_i$ , then a fraction  $1 - a$  ( $0 \leq a \leq 1$ ) of the inventory on hand spoils and is not available for future use. These assumptions imply that  $v_i(\cdot, \cdot)$  takes the form

$$v_i(y_i, D_i) = \begin{cases} a \cdot (y_i - D_i) & \text{if } y_i \geq D_i, \\ b \cdot (y_i - D_i) & \text{if } y_i \leq D_i. \end{cases}$$

Note that if  $a = 0$  we have the case of perishable goods while if  $a = 1$  we have the case of nonperishable goods. If  $b = 0$  we have the lost sales case while if  $b = 1$  we have the backlog case. In the literature these last two cases are usually discussed only where  $a = 1$ .

At the beginning of period  $i$ , the inventory manager is assumed to have observed the vector

$$H_i = (x_1, \dots, x_i, y_1, \dots, y_{i-1}, D_1, \dots, D_{i-1}),$$

representing the history of the process up to the beginning of period  $i$ . He bases his ordering decision in period  $i$  upon  $H_i$ .

An ordering policy for period  $i$  is a real valued Borel function  $\bar{Y}_i(\cdot)$  to be used as follows. At the beginning of period  $i$ , after having observed the past history  $H_i$ , the manager orders  $\bar{Y}_i(H_i) - x_i$  which is assumed to be nonnegative of course. Also let  $\bar{Y}_i = (\bar{Y}_1, \dots, \bar{Y}_n)$  denote a sequence of ordering policies for periods  $i, \dots, n$ .

Three types of costs are considered: ordering, holding, and shortage. Assume that the cost of ordering  $z$  units in period  $i$  is  $K_i \delta(z) + c_i z$ , where  $K_i \geq 0$ ,  $\delta(0) = 0$ , and  $\delta(z) = 1$  for  $z > 0$ . The cost is incurred at the time of delivery of the order. Let  $g_i(y, D_{i+\lambda})$  denote the holding and shortage cost in period  $i + \lambda$  when  $y$  is the amount of stock actually on hand after receipt of orders to be delivered before the end of period  $i + \lambda$ . We assume that  $g_i(\cdot, \cdot)$  is a real valued Borel function.

Let  $\alpha_i (\geq 0)$  be the discount factor for period  $i + \lambda$ . That is,  $\alpha_i$  is the value at the beginning of period  $i + \lambda$  of one cost unit at the beginning of period  $i + \lambda + 1$ . Let  $\beta_1 = 1$  and  $\beta_i = \prod_{j=1}^{i-1} \alpha_j$  for  $i > 1$ .

For the case  $\lambda = 0$  let

$$W_i(y, t) = c_i y + g_i(y, t) - \alpha_i c_{i+1} v_i(y, t).$$

For the case  $\lambda > 0$  let

$$W_i(y, t) = c_i y + \int_{-\infty}^{\infty} g_i(y - z, t) d\Phi_i^{\lambda}(z) - \alpha_i c_{i+1}(y - E(D_i)),$$

where  $\Phi_i^{\lambda}(\cdot)$  is the distribution of  $D_i + \dots + D_{i+\lambda-1}$ .

Now for  $\lambda \geq 0$  let

$$G_i(y) = \int_{-\infty}^{\infty} W_i(y, t) d\Phi_{i+\lambda}(t).$$

We assume that all integrals given above exist and are finite.

We suppose that each unit of stock left over after  $\lambda + n$  periods can be discarded with a return of  $c_{n+1}$ . Similarly, each unit of backlogged demand remaining after  $\lambda + n$  periods is satisfied at a cost  $c_{n+1}$ . In the literature it has often been assumed that  $c_{n+1} = 0$ .

Thus, the expected discounted cost incurred in periods  $\lambda + 1, \dots, \lambda + n$  when following the policy  $\bar{Y}_1$  in periods  $1, \dots, n$  is

$$E \left\{ \sum_{i=1}^n \beta_i \left[ K_i \delta(y_i - x_i) + c_i(y_i - x_i) + g_i \left( y_i - \sum_{j=i}^{i+\lambda-1} D_j, D_{i+\lambda} \right) \right] - \beta_{n+1} c_{n+1} \left( x_{n+1} - \sum_{i=n+1}^{n+\lambda} D_i \right) \right\}.$$

By substituting  $x_i = v_{i-1}(y_{i-1}, D_{i-1})$  into the above formula we get as in [10], [12],

$$\sum_{i=1}^n \beta_i E[K_i \delta(y_i - x_i) + G_i(y_i)] - \left[ c_1 x_1 - \beta_{n+1} c_{n+1} \sum_{i=n+1}^{n+\lambda} E(D_i) \right].$$

Since the second bracketed term is not affected by the choice of  $\bar{Y}_1$ , it is convenient to omit it from the analysis. Thus we may define the conditional expected discounted cost incurred in periods  $\lambda + i, \dots, \lambda + n$  when following  $\bar{Y}_i$  in periods  $i, \dots, n$  given the observed history  $H_i$  as

$$(1) \quad f_i(\bar{Y}_i | H_i) = \sum_{j=i}^n \beta_j E_{\pi_j}[K_j \delta(y_j - x_j) + G_j(y_j)].$$

We seek a policy  $\bar{Y}_1^* = (\bar{Y}_1^*, \dots, \bar{Y}_n^*)$ , called optimal, which satisfies

$$(2) \quad f_i(\bar{Y}_i^* | H_i) \leq f_i(\bar{Y}_i | H_i), \quad i = i, \dots, n,$$

for all  $H_i$  and  $\bar{Y}_i$ , where of course  $\bar{Y}_i^* = (\bar{Y}_i^*, \dots, \bar{Y}_n^*)$ . It is easy to show by induction on  $i$  (starting with  $i = n$ ) that if there is an optimal policy, then  $f_i(\bar{Y}_i^* | H_i)$  depends upon  $H_i$  only through  $x_i$ , so we may

write

$$(3) \quad f_i(\bar{Y}_i^* | H_i) = f_i(x_i), \quad i = 1, \dots, n,$$

where the  $f_i$  satisfy ( $f_{n+1}(x) \equiv 0$ )

$$(4) \quad f_i(x) = \inf_{y \geq x} \{K_i \delta(y - x) + G_i(y) + \alpha E f_{i+1}(v_i(y, D_i))\}$$

for  $i = 1, \dots, n$  and all  $x$  with the infimum being attained for each  $x$ . Conversely if there is a sequence of functions  $\{f_i\}$  which satisfy (4), with the infimum being attained for each  $x$ , then there exists an optimal policy  $\bar{Y}_1^*$ . Moreover,  $\bar{Y}_i^*(H_i)$  is any value of  $y$  which minimizes the expression in braces on the right side of (4) subject to  $y \geq x$  where  $x = x_i$ . Since the minimizing value of  $y$  is a function only of  $x_i$ , it follows that  $\bar{Y}_i^*(H_i) \equiv \bar{Y}_i^*(x_i)$  depends upon  $H_i$  only through  $x_i$ . For notational convenience, we define

$$(5) \quad J_i(y) = G_i(y) + \alpha E f_{i+1}(v_i(y, D_i)).$$

In what follows we shall have occasion to impose one or more of the following assumptions for each  $i$  ( $K_{n+1} \equiv 0$ ):

- (i)  $G_i(y)$  and  $v_i(y, t)$  are continuous in  $y$  for each  $t \geq \eta_i$ ;
- (ii)  $\lim_{y \rightarrow \infty} G_i(y) > \inf_y G_i(y) + \alpha K_{i+1}$ ;
- (iii)  $\lim_{y \rightarrow -\infty} G_i(y) > \inf_y G_i(y) + K_i$ ;
- (iv)  $-G_i(y)$  is unimodal in  $y$ ;
- (v)  $v_i(y, t)$  is nondecreasing in  $y$  for each  $t \geq \eta_i$ ; moreover,  $v_i(y, t)$  is bounded above in  $t$  on  $[\eta_i, \infty)$  for each fixed  $y$ ;
- (vi)  $K_i \geq \alpha K_{i+1}$ .

If (i)-(iii) hold, there are a number  $\bar{S}_i$  which minimizes  $G_i(y)$  on  $(-\infty, \infty)$  and numbers  $\bar{s}_i (\leq \bar{S}_i)$  and  $\bar{S}_i (\geq \bar{S}_i)$  such that

$$G_i(\bar{S}_i) = G_i(\bar{s}_i) + \alpha K_{i+1}$$

and

$$G_i(\bar{s}_i) = G_i(\bar{S}_i) + K_i.$$

If in addition (vi) holds, there is a number  $\bar{s}_i, \bar{s}_i \leq \bar{s} \leq \bar{S}_i$ , such that

$$G_i(\bar{s}_i) = G_i(\bar{S}_i) + (K_i - \alpha K_{i+1}).$$

**2. The optimality of the  $(s, S)$  policy.** In this section we shall show that that if (i)-(vi) hold and if

$$(vii) \quad v_i(\bar{S}_i, t) \leq \bar{S}_{i+1} \quad \text{for } t \geq \eta_i \quad \text{and } i = 1, 2, \dots, n-1,$$

then there is an optimal policy which is an  $(s, S)$  policy. By this we mean that there is a sequence  $\{(s_i, S_i)\}$  of pairs of numbers such that ( $s_i \leq S_i$ ),

$$\bar{Y}_i^*(H_i) = \begin{cases} S_i & \text{if } x_i < s_i, \\ x_i & \text{if } x_i \geq s_i, \end{cases}$$

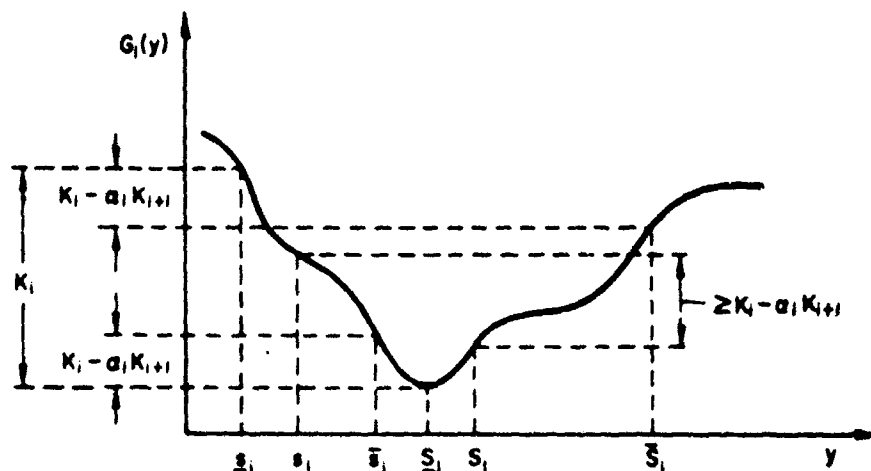


FIG. 1

for all  $i$  and  $H_i$ . Moreover the numbers satisfy

$$(6) \quad s_i \leq s_i \leq \bar{s}_i \leq \bar{S}_i \leq S_i \leq \bar{S}_i$$

and

$$(7) \quad G_i(s_i) \geq G_i(S_i) + (K_i - \alpha_i K_{i+1})$$

for all  $i$ . The bounds in (6) and (7) are depicted in Fig. 1. The inequality (7) has the interpretation that if it is optimal to order in period  $i$  with the initial inventory level  $x$ , then the reduction in the immediate expected costs due to  $G_i(\cdot)$  must be at least  $K_i - \alpha_i K_{i+1}$ . In the special case where  $\eta_i = 0$  and  $v_i(y, t) = y - t$  for all  $i$ , then (vii) reduces to the simpler form:  $S_i \leq S_{i+1}$ ,  $i = 1, \dots, n-1$ .

The first proof that the  $(s, S)$  policy is optimal under reasonably general conditions is due to Scarf [6]. (See also [13].) Scarf assumes that  $G_i(y)$  is convex in  $y$  with  $G_i(y) \rightarrow \infty$  as  $|y| \rightarrow \infty$ , that  $v_i(y, t) = y - t$ , and that (vi) holds.<sup>1</sup> These assumptions imply (i)-(vi) and are in fact quite a bit stronger. To elaborate on this point we remark that if  $W_i(y, t)$  is convex in  $y$ , then  $G_i(y)$  is convex in  $y$  for any distribution  $\Phi_{i+1}$ . However, if  $\Phi_{i+1}$  has a density  $\phi_{i+1}$  with  $\phi_{i+1}(t - \theta)$  having a monotone likelihood ratio with respect to  $\theta$ , then  $-G_i(y)$  will be unimodal under conditions where  $W_i(y, t)$  is not convex in  $y$ . Specifically,  $-G_i(y)$  is unimodal if  $W_i(y, t) = W_i^1(y - t) + W_i^2(t)$  for some functions  $W_i^1$  and  $W_i^2$  where  $-W_i^1(z)$  is unimodal in  $z$ , [3]. See [4] for a discussion of the utility of this assumption.

Although Scarf imposes stronger assumptions than (i)-(vi), he does not require that (vii) hold. Thus his results are not implied by ours nor conversely.

<sup>1</sup> Scarf also assumed that the cost functions and demand distributions do not change over time although that is not essential to his proof [8, p. 200].

As an example to illuminate the significance of (vii) suppose  $\lambda = 0$ ;  $D_1, \dots, D_n$  are identically distributed; and  $c_i = c$ ,  $\alpha_i = \alpha$ ,  $g_i(y, t) = g(y, t)$ ,  $v_i(y, t) = y - t$ , and  $\eta_i = 0$  for  $i = 1, 2, \dots, n$ . Then  $G_i(y) = G(y)$  so  $S_i = S$  for  $i = 1, 2, \dots, n - 1$ . If in addition  $c_{n+1} = c$ , then  $G_n(y) = G(y)$  also, so  $S_n = S$  and (vii) holds. On the other hand if  $c_{n+1} = 0$  (as is assumed, for example, by Scarf [6] and others), then  $G_n(y) = G(y) + \alpha[y - E(D_n)]$ . Thus if  $\alpha c > 0$ , we should ordinarily expect  $S > S_n$  in which case (viii) fails to hold for  $i = n - 1$ . Both of the above definitions of  $c_{n+1}$  are reasonable formulations of a "stationary" version of our model. It is of some interest then that the first assumption assures that (vii) holds while the second does not. Of course in infinite horizon models [1], [2], the difference between the two stationary formulations vanishes. Thus our hypotheses are actually weaker than Scarf's for the infinite horizon stationary models.

Bounds on  $s_i$  and  $S_i$  were first established by Igglehart in [1], [2] under Scarf's hypotheses, the assumption that the cost functions and demand distributions do not change over time, and  $c_{n+1} \equiv 0$ . Under the above assumptions except  $c_{n+1} = c$ , Veinott and Wagner [12] have established bounds of the form (6). In their analysis  $G_i(y) = G(y)$ , so the bounds in (6) are independent of  $i$  even though this is not true of  $s_i$  and  $S_i$ . Our present analysis shows that the bounds remain valid under the weaker hypotheses imposed here.

The principal tool of Scarf's proof is the fact that if  $J_i(y)$  is  $K_i$ -convex (see [6] for a definition), then so is  $f_i(x)$ . This method of proof fails under our hypotheses because  $J_i(y)$  need not be  $K_i$ -convex.<sup>1</sup> Our proof is based instead upon the following two lemmas which establish properties of functions satisfying (4), (5).

LEMMA 1.

$$(8) \quad f_i(x) \leq f_i(x') + K_i, \quad x \leq x', \quad i = 1, \dots, n.$$

Moreover, if (v) holds, then

$$(9) \quad J_i(y') - J_i(y) \geq G_i(y') - G_i(y) - \alpha K_{i+1}, \quad y \leq y', \\ i = 1, \dots, n.$$

*Proof.*

From (4) and (5) we have for  $x \leq x'$  that

$$f_i(x) \leq K_i + \inf_{y \geq x} J_i(y) \leq K_i + \inf_{y \geq x'} J_i(y) \leq K_i + f_i(x'),$$

which establishes (8).

<sup>1</sup>As an illustration, if  $K_i = K < 1$ ,  $G_i(y) = G(y) = \min(1, |y|)$ ,  $v_i \geq 0$ ,  $s_i(y, t) = y - t$ , and  $\alpha_i = \alpha \leq 1$ , then (v) holds. However,  $J_i(y) = G(y)$  which is not  $K_i$ -convex.

For  $y \leq y'$ , we have from (v) that  $v_i(y, D_i) \leq v_i(y', D_i)$ . Thus from (4), (5), and (8) we get

$$\begin{aligned} J_i(y') - J_i(y) &= G_i(y') - G_i(y) + \alpha_i E[f_{i+1}(v_i(y', D_i)) - f_{i+1}(v_i(y, D_i))] \\ &\geq G_i(y') - G_i(y) - \alpha_i K_{i+1}, \end{aligned}$$

which completes the proof.

The proofs of (8) and (9) are purely analytic. An alternative proof of (8) may be constructed by using the following argument which may be made rigorous. If the initial inventory on hand and on order at the beginning of period  $i$  is  $x$  but one orders in each period  $j$  ( $\geq i$ ) so as to bring the inventory level after ordering to the level which would be optimal if the initial inventory level in period  $i$  were instead  $x'$  ( $\geq x$ ), then the associated expected discounted cost would not exceed  $f_i(x') + K_i$ . But since the policy just described cannot be better than the optimal policy, (8) must hold. A similar kind of argument may be used to establish (9).

LEMMA 2. If (v) holds, if  $\{a_j\}$  is a sequence of numbers for which  $v_j(a_j, t) \leq a_{j+1}$  for  $t \geq \eta_j$  and  $j \geq i$ , and if  $G_j(y)$  is nonincreasing in  $y$  on  $(-\infty, a_j]$  for  $j \geq i$ , then

$$(10) \quad J_j(y') - J_j(y) \leq G_j(y') - G_j(y) \leq 0, \quad y \leq y' \leq a_j,$$

and

$$(11) \quad f_j(x') - f_j(x) \leq 0, \quad x \leq x' \leq a_j,$$

for  $j \geq i$ .

*Proof.* The proof is by induction on  $j$ . Suppose (10), (11) hold for  $j+1$  ( $> i$ ). By (v),  $v_j(y, D_j) \leq v_j(y', D_j) \leq v_j(a_j, D_j) \leq a_{j+1}$ . Hence using (11) for  $j+1$  we get

$$\begin{aligned} J_j(y') - J_j(y) &= G_j(y') - G_j(y) + \alpha_j E[f_{j+1}(v_j(y', D_j)) - f_{j+1}(v_j(y, D_j))] \\ &\leq G_j(y') - G_j(y) \leq 0, \end{aligned}$$

which proves (10) for  $j$ .

It follows from (10) for the integer  $j$  that

$$\begin{aligned} f_j(x) &= \min \{J_j(x), K_j + \inf_{y \geq x} J_j(y)\} \\ &\geq \min \{J_j(x'), K_j + \inf_{y \geq x'} J_j(y)\} = f_j(x'), \end{aligned}$$

which proves (11) for the integer  $j$ . The same arguments suffice to establish (10), (11) for  $j = n$  which starts the induction and completes the proof.

The proofs of (10) and (11) are purely analytic. An alternative proof of (11) may be devised by using the following argument which can be made rigorous. Suppose the initial inventory level in period  $j$  is  $x'$ . Suppose also that one orders so as to bring the initial inventory level after ordering in each period  $k$  ( $\geq j$ ) as close as possible to the level which would be optimal if the initial inventory level in period  $j$  were  $x$  ( $\leq x'$ ). This policy incurs expected discounted costs which are no greater than  $f_j(x)$ . But the policy also must incur expected discounted costs at least as large as  $f_j(x')$ . Combining these remarks proves (11). The inequality (10) can be justified in a similar way.

**THEOREM 1.** *If (i)–(vii) hold, there exists an optimal policy which is an  $(s, S)$  policy. Moreover, the parameters of that policy satisfy (6), (7).*

*Proof.* The proof is constructive and proceeds in several steps. To begin with suppose  $J_i(y)$  is continuous in  $y$ .

(a)  $J_i(y)$  is nonincreasing on  $(-\infty, \underline{S}_i]$ .

To see this recall from (iv) that  $G_j(y)$  is nonincreasing in  $y$  on  $(-\infty, \underline{S}_j]$  for  $j \geq i$ . Thus by (vii) and Lemma 2, (a) holds.

Since  $J_i(y)$  is continuous, there is an  $S_i$  which minimizes  $J_i(y)$  on  $[\underline{S}_i, \bar{S}_i]$ . Thus  $S_i$  satisfies (6). Moreover,

(b)  $\min_y J_i(y) = J_i(S_i)$ .

To see this observe from (a) that  $S_i$  minimizes  $J_i(y)$  on  $(-\infty, \bar{S}_i]$ . Also by Lemma 1, (iv), and the definition of  $\bar{S}_i$ , we have for  $y > \bar{S}_i$  that

$$\begin{aligned} J_i(y) - J_i(\underline{S}_i) &\geq G_i(y) - G_i(\underline{S}_i) - \alpha K_{i+1} \\ &\geq G_i(\bar{S}_i) - G_i(\underline{S}_i) - \alpha K_{i+1} = 0. \end{aligned}$$

Thus (b) holds.

(c) There exists a number  $s_i$  satisfying (6), (7) and

$$(12) \quad J_i(S_i) + K_i - J_i(s_i) = 0.$$

In order to prove this assertion we observe from Lemma 2, (b), and the definitions of  $\underline{S}_i$  and  $\underline{s}_i$  that

$$\begin{aligned} (13) \quad J_i(S_i) + K_i - J_i(\underline{s}_i) &\leq J_i(\underline{S}_i) + K_i - J_i(\underline{s}_i) \\ &\leq G_i(\underline{S}_i) + K_i - G_i(\underline{s}_i) = 0. \end{aligned}$$

On the other hand by Lemma 1 and the definitions of  $\underline{S}_i$  and  $\bar{S}_i$  we have

$$\begin{aligned} (14) \quad J_i(S_i) + K_i - J_i(\bar{S}_i) &\geq G_i(S_i) - G_i(\bar{S}_i) + K_i - \alpha K_{i+1} \\ &\geq G_i(\underline{S}_i) - G_i(\bar{S}_i) + K_i - \alpha K_{i+1} = 0. \end{aligned}$$

From (13), (14), and the continuity of  $J_i(y)$ , it follows that there is an  $s_i$  satisfying (6) and (12). Moreover (7) holds also since by Lemma 1

and (12) we have

$$0 = J_i(S_i) + K_i - J_i(s_i) \geq G_i(S_i) - G_i(s_i) + K_i - \alpha_i K_{i+1}$$

which completes the proof of (c).

(d) The value of  $y$  which minimizes the right side of (4) is determined by

$$y = \begin{cases} S_i & \text{if } x < s_i, \\ x & \text{if } x \geq s_i. \end{cases}$$

To prove (d), observe from (a), (b), and (c) that for  $x < s_i$ ,

$$J_i(x) \geq J_i(s_i) = J_i(S_i) + K_i = \min_y J_i(y) + K_i,$$

so  $y = S_i$  minimizes the right side of (4). Now for  $s_i \leq x \leq S_i$ , the same arguments give

$$J_i(x) \leq J_i(s_i) = \min_y J_i(y) + K_i,$$

so  $y = x$  minimizes the right side of (4). Finally for  $S_i < x < y$ , we have from Lemma 1, (iv), and (vi) that

$$J_i(y) + K_i - J_i(x) \geq G_i(y) - G_i(x) + K_i - \alpha_i K_{i+1} \geq 0,$$

so  $y = x$  minimizes the right side of (4). This completes the proof of (d).

It remains only to verify our assumption that

(e)  $J_i(y)$  is continuous in  $y$ .

We prove (e) by induction on  $i$ . The assertion is trivial for  $i = n$  since  $J_n(y) = G_n(y)$ . Suppose now (e) holds for the integer  $i + 1$ . Then by (d),

$$(15) \quad f_{i+1}(x) = \begin{cases} K_{i+1} + J_{i+1}(S_{i+1}) & \text{if } x < s_{i+1}, \\ J_{i+1}(x) & \text{if } x \geq s_{i+1}. \end{cases}$$

Since  $J_{i+1}(y)$  is continuous and (12) holds for  $i+1$ ,  $f_{i+1}(x)$  is evidently continuous. Since  $G_i(y)$  is continuous by (i),  $J_i(y)$  will be continuous if

$$Ef_{i+1}(v_i(y, D_i)) \equiv q(y)$$

is continuous in  $y$  on any arbitrary interval,  $[a, b]$ , say. Since by (i) and the continuity of  $f_{i+1}$ , the composite function  $f_{i+1}(v_i(y, t))$  is continuous in  $y$ ,  $q(y)$  will be continuous on  $[a, b]$  if the composite function is uniformly bounded for  $y \in [a, b]$  and all  $t$  ( $\geq \eta_i$ ) by virtue of the dominated convergence theorem. We now show that there is a number  $B$  such that

$$(16) \quad J_{i+1}(S_{i+1}) \leq f_{i+1}(v_i(y, t)) \leq f_{i+1}(B) + K_{i+1}$$

for all  $t$  ( $\geq \eta_i$ ) and  $y \in [a, b]$ , which gives the desired bounds. The left-hand inequality follows from (15) and (b). Since by (v),  $v_i(y, t) \leq v_i(b, t) \leq B$  for some  $B$ , the right-hand inequality follows from Lemma 1.

The proof is now complete since we have constructed a solution to (4) with the infimum being attained for each  $x$ .

As we have remarked before, if Scarf's hypotheses ( $G_i(y)$  convex,  $G_i(y) \rightarrow \infty$  as  $|y| \rightarrow \infty$ ,  $v_i(y, t) = y - t$ , and (vi)) are substituted for (i)-(vii), then there exists an optimal policy which is an  $(s, S)$  policy. However, the lower bounds in (6) for  $s_i$  and  $S_i$  are no longer valid when (vii) fails to hold. The reason for this is clear upon reflection. For example, suppose  $D_{n-1} \geq 0$  and  $P(D_{n-1} < \underline{S}_{n-1} - \underline{S}_n) > 0$  so (vii) does not hold. In this event it is apparent from (1) that one would not want to order up to  $\underline{S}_{n-1}$  (or more) in period  $n - 1$  if  $G_n(y)$  increased sufficiently rapidly on the interval  $[\underline{S}_n, \underline{S}_{n-1}]$ . The reason for this is, of course, that the relatively low expected costs in period  $n - 1$  would be more than offset by the extremely high expected costs in period  $n$ . For a concrete illustration see footnote 4 below.

Let

$$\underline{S}_i = \begin{cases} \underline{S}_n & \text{if } i = n, \\ \min(\underline{S}_i, \underline{S}_{i+1} + \eta_i) & \text{if } i = 1, 2, \dots, n - 1. \end{cases}$$

Let  $\underline{s}_i$  ( $\leq \min(\underline{s}_i, \underline{S}_i)$ ) be chosen so that

$$G_i(\underline{s}_i) = G_i(\underline{S}_i) + K_i.$$

**THEOREM 2.** *Under Scarf's hypotheses, there is an optimal policy  $\{(s_i, S_i)\}$  which satisfies (7) and*

$$(6') \quad \underline{s}_i \leq s_i \leq \bar{s}_i \quad \text{and} \quad \underline{S}_i \leq S_i \leq \bar{S}_i, \quad i = 1, 2, \dots, n.$$

*Proof.* We only sketch the proof, leaving the details to the reader. The upper bounds on  $s_i$  and  $S_i$  and the inequality (7) are established by applying Lemma 1 in exactly the same way as in Theorem 1. The lower bounds on  $s_i$  and  $S_i$  may be established by applying (10) with  $a_j = \underline{S}_j$  for all  $j$  in a manner similar to that employed in proving Theorem 1.

We remark that if (vii) holds then  $\underline{S}_i = \underline{S}_i$  and  $\underline{s}_i = s_i$  for all  $i$  so that (6') reduces to (6).

**3. Planning horizons and special cases.** The next result tells us that if  $S_k$  is sufficiently small in comparison with  $s_{k+1}$ , then  $(\underline{s}_k, \underline{S}_k)$  is optimal for period  $k$ . Observe that this is the policy that is optimal for period  $k$  when considered by itself or as the final period of a  $k$ -period model. Moreover, if  $S_1, \dots, S_k$  are sufficiently small in comparison with  $s_{k+1}$ , an optimal policy for periods  $i, \dots, k$  may be determined without evaluating  $f_{k+1}(x)$  for any  $x$ . In this sense, period  $k$  is a planning horizon. The actual calculations are carried out using (4) recursively where  $f_{k+1}(x) = 0$  for all  $x$ . The theorem generalizes some results in [9] to the case where there is a setup cost for placing orders.

**THEOREM 3.** *Suppose (i)-(vii) or Scarf's hypotheses hold, and that  $\{(s_j, S_j)\}$  is an optimal policy.*

(a) If  $\{a_j\}$  is a collection of numbers for which

$$(17)^4 \quad S_k \leq a_k,$$

$$(18) \quad a_{k+1} \leq s_{k+1},$$

$$(19) \quad v_j(a_j, t) \leq a_{j+1} \text{ for } t \geq \eta_j, \quad \text{and}$$

$$(20) \quad S_j \leq a_j$$

for  $j = k$ , then  $(s_k, S_k)$  is optimal for period  $k$ .

(b) If (18)–(20) hold for  $j = i, i+1, \dots, k$ , then one optimal policy for periods  $i, i+1, \dots, k$ , is independent of  $f_{k+1}(\cdot)$ .

*Proof.* We begin by proving part (a). From Theorem 1, Scarf's theorem, and (18),

$$(21) \quad f_{k+1}(x) = K_{k+1} + J_{k+1}(S_{k+1}) \equiv Q, \quad x \leq a_{k+1}.$$

It follows from (v) and (19) that for  $t \geq \eta_k$  and  $y \leq a_k$ ,  $v_k(y, t) \leq v_k(a_k, t) \leq a_{k+1}$ . Combining this remark and (21) we get

$$(22) \quad J_k(y) = G_k(y) + \alpha_k E f_{k+1}(v_k(y, D_k)) = G_k(y) + \alpha_k Q, \quad y \leq a_k.$$

Now by (20),  $J_k(y)$  achieves its minimum on  $(-\infty, \infty)$  in  $(-\infty, a_k]$ . Since this is so it follows from (22) and (17) that  $S_k$  minimizes  $J_k(y)$  on  $(-\infty, \infty)$ . Moreover, again by (22) we have

$$J_k(s_k) = G_k(s_k) + \alpha_k Q = K_k + G_k(S_k) + \alpha_k Q = K_k + J_k(S_k).$$

Hence, by Theorem 1 and Scarf's theorem,  $(s_k, S_k)$  is optimal for period  $k$ , which establishes part (a).

In order to prove part (b) we observe from Theorem 1, Scarf's theorem, and (20) that the optimal policy in period  $j$ ,  $i \leq j \leq k$ , can be determined provided only that we can evaluate  $J_j(y)$  for  $y \leq a_j$ . Now from (v), (19), and (20) it follows easily by induction on  $j$  that  $J_j(y)$  may be evaluated for  $y \leq a_j$  without evaluating  $f_{k+1}(x)$  for  $x > a_{k+1}$ , i.e., for  $y \leq a_j$ ,  $J_j(y)$  depends upon  $f_{k+1}$  only through the constant  $Q$ . (We have already shown this for  $j = k$  which starts the induction.)

We may exhibit the dependence of  $J_j(y)$  upon  $Q$  by writing  $J_j^0(y)$ . It is easy to show by induction on  $j$  that

$$(23) \quad J_j^0(y) = J_j^0(y) + Q \prod_{i=j}^k \alpha_i, \quad j = i, i+1, \dots, k.$$

<sup>4</sup> The hypothesis (17) is easily seen to be satisfied if (18)–(20) hold for  $j = k$  and either (1) (i)–(vii) hold or (2) Scarf's assumptions are fulfilled and  $K_{k+1} > 0$  or  $S_k < a_k$ . The following example shows that (17) cannot be dispensed with under Scarf's hypotheses. Assume  $n = k+1 = 2$ ,  $G_1(y) = |y-2|$ ,  $G_2(y) = 2|y-1|$ ,  $K_1 = K_2 = 0$ ,  $P(D_1 = 0) = P(D_2 = 0) = 1$ , and  $\eta_1 = \eta_2 = 0$ . Then  $s_1 = S_1 = s_2 = S_2 = 1$  and  $s_1 = S_1 = 2$ . Now let  $a_1 = a_2 = 1$ . Then (18)–(20) hold, but  $(s_1, S_1)$  is not optimal for period 1.

(Again we have already done this for  $j = k$  which starts the induction.) Thus if a policy is optimal for period  $j$ ,  $i \leq j \leq k$ , for some  $Q$ , that same policy is optimal for all  $Q$ . Hence if we assume  $Q = 0$  and determine the optimal policy for periods  $i, i + 1, \dots, k$  in the usual way (taking account of (20)), that policy is optimal for the original problem where (21) holds. We have thus shown that the optimal policy in periods  $i, i + 1, \dots, k$  is independent of  $f_{k+1}(\cdot)$  as required.

As an illustration of the application of Theorem 3(a), we have the following result.

**COROLLARY 1.** *If (i)-(vii) or Scarf's hypotheses hold, and if*

$$(24) \quad v_k(\bar{S}_k, t) \leq g_{k+1} \quad \text{for } t \geq \eta_k,$$

*then  $(g_k, \bar{S}_k)$  is optimal for period  $k$ .*

*Proof.* Let  $a_k = \bar{S}_k$  and  $a_{k+1} = g_{k+1}$ . Then apply Theorem 3(a).

We remark that if (24) holds, then by (v) and the definitions of  $\bar{S}_k$ ,  $\bar{S}_k, g_{k+1}, \bar{S}_{k+1}$  (recall  $g_{k+1} = \bar{S}_{k+1}$  if (i)-(vii) hold),

$$v_k(\bar{S}_k, t) \leq v_k(\bar{S}_k, t) \leq g_{k+1} \leq \bar{S}_{k+1},$$

so (vii) holds for the integer  $k$ . It follows therefore that if (24) holds for all  $k$ , then (vii) necessarily holds also. In this event we can replace  $g_{k+1}$  in (24) by  $\bar{S}_{k+1}$ .

**Example 1.** Suppose unsatisfied demand is backlogged in period  $k$  so that  $v_k(y, t) = y - t$ . Then (24) reduces to  $\bar{S}_k - g_{k+1} \leq \eta_k$ . Thus if the minimal demand in period  $k$  is at least  $\bar{S}_k - g_{k+1}$ , then  $(g_k, \bar{S}_k)$  is optimal for period  $k$  by Corollary 1. A special case of this result is established in [12, p. 545] for the stationary case.

The next example illustrates the application of Theorem 3(b).

**Example 2.** Suppose unsatisfied demands are backlogged so that  $v_j(y, t) = y - t$ . Also let  $a_{k+1} = g_{k+1}$  and let  $\{a_j\}$  be defined recursively by  $a_j - \eta_j = a_{j+1}$  for  $j \leq k$ . Thus  $a_i = g_{k+1} + \sum_{j=i}^k \eta_j$  for  $i \leq k$ . Now let  $i$  be an integer for which

$$(25) \quad \bar{S}_j \leq g_{k+1} + \sum_{t=j}^k \eta_t, \quad j = i, i + 1, \dots, k.$$

Then the hypotheses of Theorem 3(b) are evidently satisfied. In particular, if  $\eta_t = 0$  for  $i \leq t \leq k$ , then the optimal policy in periods  $i, i + 1, \dots, k$  may be determined without evaluating  $f_{k+1}(\cdot)$  if  $\bar{S}_j \leq g_{k+1}$  for  $i \leq j \leq k$ .

We remark that if (i)-(vii) hold and if  $K_i = 0$  for all  $i$ , then we choose  $g_i, \bar{S}_i$ , and  $\bar{S}_i$  equal to  $\bar{S}_i$  so  $s_i = \bar{S}_i = \bar{S}_i$  for all  $i$  by Theorem 1. In this event the hypothesis (vii) is equivalent to the hypothesis that (24) holds for all  $k$ . This observation together with Corollary 1 establishes the next result which is known from [10], [11].

**COROLLARY 2.** *If (i)–(vii) hold and if  $K_i = 0$  for all  $i$ , then  $\{(S_i, s_i)\}$  is an optimal policy.*

**4. Applications and extensions.** In this section we discuss some applications and extensions of our results.

*Applications of the basic lemmas.* Lemmas 1 and 2 are useful in establishing bounds on the ordering regions and order quantities even where  $-G_i(y)$  is not unimodal. We shall illustrate this point under the assumption that  $K_i = 0$  for all  $i$ , leaving the other case to the reader. Suppose  $G_i(y)$  appears as in Fig. 2. The domain of  $G_i(y)$  is divided into six regions labeled 1, 2,  $\dots$ , 6. If period  $i$  were considered by itself, and if the initial inventory in period  $i$  fell in an odd-numbered region, it would be optimal to order to the upper bound of that region, viz., to  $U_1$ ,  $U_3$ , or  $U_5$  as appropriate. If the initial inventory in period  $i$  fell in an even-numbered region, no order should be placed. If  $i < n$ , then the above policy need not be optimal for the  $n$ -period model. However, suppose  $y$  lies in an even-numbered region. Then by Lemma 1,

$$(26) \quad J_i(y') - J_i(y) \geq G_i(y') - G_i(y) \geq 0$$

for all  $y' \geq y$  provided (v) holds so it is optimal not to order in period  $i$  for the  $n$ -period model. Notice also that the above inequality tells us that if the initial inventory level in period  $i$  is  $a$  and if it is optimal to order, then the inventory level after ordering must lie in the interval  $[b, c]$ .

If (v) holds, if  $v_j(U_1, t) \leq U_1$  for  $t \geq \tau_j$ , and if  $G_j(y)$  is nonincreasing in  $y$  on  $(-\infty, U_1]$  for  $j \geq i$ , then by Lemma 2,  $J_i(y)$  is minimized on  $(-\infty, U_1]$  at  $y = U_1$ . Similarly from (26),  $J_i(y)$  is minimized on  $[U_1, \infty)$  at  $y = U_1$ . Combining these remarks we see that  $J_i(y)$  is minimized on  $(-\infty, \infty)$  at  $y = U_1$ . Hence, in region 1 it is optimal to order up to  $U_1$ .

*Variation of the bounds over time.* It is of interest to determine how  $s_i$

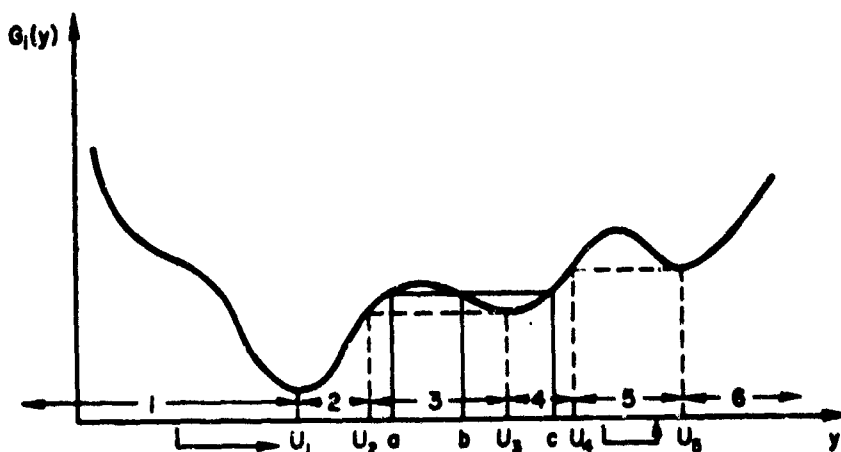


FIG. 2

and  $S_i$  vary over time in relation to the variation of the cost functions and demand distributions. Although this appears to be quite difficult, we can instead examine how the bounds on  $s_i$  and  $S_i$  vary over time. Such studies are of interest in their own right and because they provide us with a tool for determining conditions under which the hypotheses (vii), (17)–(20), (24), (25) of our several results hold. Throughout this subsection we shall assume for simplicity that sufficient regularity conditions are imposed to permit differentiation and interchange of differentiation with integration where required.

As a preliminary we record several lemmas from [11]. Let  $I$  be a subset of the integers  $1, \dots, n$ .

LEMMA 3. If  $\partial W_i(y, t)/\partial y$  is nonincreasing in  $t \geq \eta_i$  and  $i \in I$  for each  $y$ , and if  $\Phi_i(t) \geq \Phi_j(t)$  for all  $t$  and  $i, j \in I, i < j$ , then

$$(27) \quad G'_i(y) \geq G'_j(y), \quad i, j \in I, \quad i < j, \quad \text{and all } y.$$

LEMMA 4. If  $W_i(y, t) = W_i^1(y - t) + W_i^2(t)$  for some functions  $W_i^1, W_i^2$ , if  $dW_i^1(z)/dz$  is nonincreasing in  $i \in I$  and nondecreasing in  $z$ , and if  $\Phi_i(t) \geq \Phi_j(t - b_{ij})$  for all  $t$  and  $i, j \in I, i < j$ , and some numbers  $b_{ij}$ , then

$$(28) \quad G'_i(y) \geq G'_j(y - b_{ij}), \quad i, j \in I, \quad i < j, \quad \text{and all } y.$$

LEMMA 5. If  $\lambda = 0$ , if  $W_i(y, t) = W^1(y - t) + W_i^2(t)$  for some functions  $W^1, W_i^2$ , and if  $\Phi_i(t) = \Phi(t - \eta_i)$  for all  $t$  and  $i \in I$  for some distribution  $\Phi$ , then

$$(29) \quad G_i(y) = G(y - \eta_i) + Q_i, \quad i \in I, \quad \text{and all } y,$$

where  $Q_i$  is a constant and  $G(y) = \int_{-\infty}^{\infty} W(y - t) d\Phi(t)$ .

In the remainder of this subsection we assume for simplicity that  $K_i = K$  and  $\alpha_i = \alpha$  ( $\leq 1$ ) for all  $i$ , and  $\lambda = 0$ . Our methods can be applied without these hypotheses, but not without expanding the exposition. See in particular [11] for conditions under which the hypotheses of Lemmas 3 and 4 are satisfied when  $\lambda > 0$ . There is also an analog of Lemma 5 when  $\lambda > 0$ .

Let  $\mathcal{S}_i = (s_i, \bar{s}_i, S_i, \bar{S}_i)$ ,  $B_{ij} = (b_{ij}, \bar{b}_{ij}, b_{ij}, \bar{b}_{ij})$ , and  $H_i = (\eta_i, \bar{\eta}_i, \eta_i, \bar{\eta}_i)$ . Let  $\mathcal{S} = (s, \bar{s}, S, \bar{S})$ , where  $s, \bar{s}, S, \bar{S}$  are defined for the function  $G(\cdot)$  (see Lemma 5) in the usual way. For definiteness where  $\mathcal{S}_i$  is not uniquely defined we choose it as follows. First pick the smallest possible  $S_i$ . Then pick the smallest  $s_i, \bar{s}_i$ , and  $\bar{S}_i$ . Do the same for  $\mathcal{S}$ . The following theorem, which is an easy consequence of Lemmas 3–5, describes how  $\mathcal{S}_i$  varies over time in relation to the variation of  $W_i$  and  $\Phi_i$  (as reflected in  $G_i$ ) over time.

**THEOREM 4.** Suppose (i)-(iv) hold.

- (a) If (27) holds, then  $s_i \leq s_j$  for  $i, j \in I, i < j$ .
- (b) If (28) holds, then  $s_i - B_{i,j} \leq s_j$  for  $i, j \in I, i < j$ .
- (c) If (29) holds, then  $s_i - H_i = s$  for  $i \in I$ .

*Satisfying the hypotheses of the main results.* In this subsection we use Theorem 4 to give conditions under which the important hypotheses (vii) and (24) of Theorem 1 and Corollary 1 respectively are satisfied. We begin by giving conditions under which (vii) holds.

It will be convenient in what follows to assume that there is an extended real number  $\theta$  such that

$$(30) \quad v_i(y, t) \leq \max(\theta, y - \eta_i), \quad \text{for } t \geq \eta_i, \quad \text{all } i, \quad \text{and all } y.$$

As an example, if  $\eta_i \geq 0$  and unsatisfied demands are backlogged so  $v_i(y, t) = y - t$ , then (30) holds with  $\theta \geq -\infty$ . Alternatively if  $\eta_i \geq 0$  and if unsatisfied demands are lost, so  $v_i(y, t) = \max(y - t, 0)$ , then (30) holds with  $\theta \geq 0$ . In applications  $\theta$  should be chosen as small as possible. Thus  $\theta = -\infty$  in the backlog case and  $\theta = 0$  in the lost sales case.

The following result is a simple consequence of Theorem 4.

**COROLLARY 3.** Suppose (i)-(iv) and (30) hold,  $I = \{1, 2, \dots, n\}$ , and  $\underline{s}_i \geq \theta$  for all  $i > 1$ .

- (a) If (27) holds and  $\eta_i \geq 0$  for all  $i < n$ , then (vii) holds.
- (b) If (28) holds and  $\eta_i - b_{i,i+1} \geq 0$  for all  $i < n$ , then (vii) holds.
- (c) If (29) holds and  $\eta_i \geq 0$  for all  $i \leq n$ , then (vii) holds.

The next corollary gives a condition ensuring that the hypothesis (24) of Corollary 1 holds.

**COROLLARY 4.** If (i)-(iv), (vii), (30) hold, if (28) holds with  $I = \{k, k+1\}$ , if  $\underline{s}_{k+1} \geq \theta$ , and if

$$(31) \quad \bar{s}_k - \underline{s}_k \leq \eta_k - b_{k,k+1},$$

then (24) holds.

*Proof.*  $v_k(\bar{s}_k, t) \leq \max(\theta, \bar{s}_k - \eta_k) \leq \max(\theta, \underline{s}_k - b_{k,k+1}) \leq \max(\theta, \underline{s}_{k+1}) = \underline{s}_{k+1} = \bar{s}_{k+1}$ .

*Stationary infinite horizon models.* This paper is primarily concerned with a finite horizon model. If the model is stationary, i.e.,  $G_i, K_i, \alpha_i, \Phi_i$  are independent of  $i$ , then it is convenient to consider an infinite period version of the model. In this case fairly obvious modifications of Iglehart's results and proofs [1], [2] (see also [12, pp. 530-531]) for  $0 \leq \alpha \leq 1$  show that if (i)-(vii) hold, there is an optimal  $(s, S)$  policy with the optimal choice of parameters being independent of time and satisfying (6), (7). Methods for computing these parameters are discussed in [8] and [12].

*Restrictions on inventory levels.* In some applications it may be desirable to limit the choice of the inventory  $y$ , on hand and on order after ordering

in period  $i$  ( $= 1, 2, \dots, n$ ) to an interval  $[y_i, \bar{y}_i]$  say. The upper bound  $\bar{y}_i$  might reflect limitations on storage space while the lower bound  $y_i$  could reflect a desire to limit the size of the backlogged demand. As a specific illustration, suppose demands occur over only the first  $n - 1$  periods, so  $D_i = 0, i \geq n$ . Then we may wish to require that no unsatisfied demand exist at the end of period  $n + \lambda$ .<sup>5</sup> This may be accomplished by setting  $y_n = 0$  so  $y_n \geq 0$ . This implies  $y_{n+\lambda} \geq 0$  if  $v_i(y, 0) \geq 0$  for  $y \geq 0$  and  $n \leq i$ .

In other applications it is natural to suppose that the demands are integers. Of course this restriction is already allowed in our formulation. However, in such cases it is usually necessary to impose the additional restriction that the order quantities and stock levels be integers. We shall now generalize our original model and results to provide for such integer restrictions and for bounds on the stock levels.

Let  $Y_i$  denote the nonempty set of admissible stock levels  $y_i$  on hand and on order after ordering in period  $i$ . Let  $y_i = \inf Y_i$  and  $\bar{y}_i = \sup Y_i$ . Let  $\mathcal{D}_i$  denote the (Borel) set of possible values of the demand  $D_i$  in period  $i$ . Let  $X_{i+1}$  denote the nonempty set of possible values of the stock on hand and on order before ordering in period  $i + 1$ . We naturally impose the consistency condition that  $v_i(y, t) \in X_{i+1}$  for all  $y \in Y_i$  and  $t \in \mathcal{D}_i$ . In addition we suppose that if the stock on hand and on order before ordering in period  $i$  is at least  $y_i$  in period  $i$ , then it is possible *not* to order in period  $i$ . Formally, we assume that  $x \in X_i$  and  $y_i \leq x$  imply  $x \in Y_i, i = 1, 2, \dots, n$ , where  $X_1 = \{x_1\}$ . We also suppose that the domains of  $f_i(\cdot), G_i(\cdot), J_i(\cdot)$ , and  $v_i(\cdot, \cdot)$  are respectively  $X_i, Y_i, Y_i$ , and  $Y_i \times \mathcal{D}_i$ . Moreover, we shall replace (i)–(iii) respectively by:

- (i') (i) holds and  $Y_i$  is closed;
- (ii') either (ii) holds or  $\bar{y}_i < \infty$ ;
- (iii') either (iii) holds or  $-\infty < y_i$ .

If (i') holds, we may define

$$G_i^+(y) = G_i(\inf \{z \mid z > y, z \in Y_i\})$$

for  $y < \bar{y}_i$  and  $G_i^+(\bar{y}_i) = \infty$  if  $\bar{y}_i < \infty$ . Also

$$G_i^-(y) = G_i(\sup \{z \mid z < y, z \in Y_i\})$$

for  $y_i < y$  and  $G_i^-(y_i) = \infty$  if  $-\infty < y_i$ . For example, if  $Y_i = (-\infty, \infty)$ , then  $G_i^-(y) = G_i(y) = G_i^+(y)$ , while if  $Y_i$  is the set of integers, then  $G_i^-(y) = G_i(y - 1)$  and  $G_i^+(y) = G_i(y + 1)$  for  $y \in Y_i$ .

If (i')–(iii') hold, there are a number  $\bar{S}_i \in Y_i$  that minimizes  $G_i(\cdot)$  on  $Y_i$  and numbers  $\underline{s}_i$  ( $\leq \bar{S}_i$ ) and  $\bar{S}_i$  ( $\geq \bar{S}_i$ ) such that

$$G_i(\bar{S}_i) \leq G_i(\underline{S}_i) + \alpha_i K_{i+1} \leq G_i^+(\bar{S}_i)$$

<sup>5</sup> I am indebted to G. Lieberman for a discussion of this application.

and

$$G_i(s_i) \leq G_i(S_i) + K_i \leq G_i^-(s_i).$$

If in addition (vi) holds, there is a number  $\bar{s}_i$ ,  $s_i \leq \bar{s}_i \leq S_i$ , such that

$$G_i(\bar{s}_i) \leq G_i(S_i) + K_i - \alpha_i K_{i+1} \leq G_i^-(\bar{s}_i).$$

It is easy to check that the statements and proofs of Lemmas 1 and 2 remain valid in our new setup. Also Theorem 1 holds provided we replace (i)-(iii) by (i')-(iii') and replace (7) by

$$(7') \quad G_i^-(s_i) \geq G_i(S_i) + (K_i - \alpha_i K_{i+1}).$$

Only obvious modifications of the proof of Theorem 1 are required.

#### REFERENCES

- [1] D. IGLEHART, *Dynamic programming and stationary analysis of inventory problems*, [7, Chap. 1].
- [2] ———, *Optimality of (s, S) policies in the infinite horizon dynamic inventory problem*, *Management Sci.*, 9 (1963), pp. 259-267.
- [3] S. KARLIN, *Polya type distributions, II*, *Ann. Math. Statist.*, 28 (1957), p. 202.
- [4] ———, *One stage inventory models with uncertainty*, *Studies in the Mathematical Theory of Inventory and Production*, K. Arrow, S. Karlin, and H. Scarf, eds., Stanford University Press, Stanford, 1968, Chap. 8.
- [5] ———, *Dynamic inventory policy with varying stochastic demands*, *Management Sci.*, 6 (1960), pp. 231-258.
- [6] H. SCARF, *The optimality of (S, s) policies in the dynamic inventory problem*, *Mathematical Methods in the Social Sciences*, K. Arrow, S. Karlin, and P. Suppes, eds., Stanford University Press, Stanford, 1960, Chap. 13.
- [7] H. SCARF, D. GILFORD, AND M. SHELLY, eds., *Multistage Inventory Models and Techniques*, Stanford University Press, Stanford, 1963.
- [8] H. SCARF, *A survey of analytic techniques in inventory theory*, [7, Chap. 7].
- [9] A. VEINOTT, JR., *Optimal stockage policies with nonstationary stochastic demands*, [7, Chap. 4].
- [10] ———, *Optimal policy for a multiproduct, dynamic nonstationary inventory problem*, *Management Sci.*, 12 (1966), pp. 206-222.
- [11] ———, *Optimal policy in a dynamic, single product, nonstationary inventory model with several demand classes*, *Operations Res.*, 13 (1965), pp. 761-778.
- [12] A. VEINOTT, JR., AND H. WAGNER, *Computing optimal (s, S) inventory policies*, *Management Sci.*, 11 (1965), pp. 525-532.
- [13] E. ZABEL, *A note on the optimality of (S, s) policies in inventory theory*, *Ibid.*, 9 (1963), pp. 123-125.

**DIFFUSION APPROXIMATIONS  
IN APPLIED PROBABILITY**

by

**DONALD L. IGLEHART**

at the

**American Mathematical Society Summer Seminar**

on the

**Mathematics of the Decision Sciences**

**Stanford University**

**July - August 1967**

# DIFFUSION APPROXIMATIONS IN APPLIED PROBABILITY<sup>1</sup>

by

Donald L. Iglehart<sup>2</sup>  
Cornell University

## 1. Introduction and Summary

For many problems in applied probability it is difficult to obtain explicit expressions for the distributions of random quantities of interest. In some problems however, approximations to these distributions can be obtained from limit theorems as a particular physical parameter approaches a limit. These approximations are similar in spirit to the normal approximation, for sums of independent random variables, resulting from the central limit theorem. Our purpose in this expository paper is to sketch the general approach to these limit theorems and approximations and to mention a number of techniques which have proved to be useful in various probabilistic models.

The starting point for us is a given sequence of stochastic processes  $(X_n(k): k = 0, 1, \dots)$  for  $n = 1, 2, \dots$  with each process defined on its own probability space  $(\Omega_n, \mathcal{F}_n, P_n)$ . We have indicated a discrete time parameter for these processes, however, this is not crucial and, in fact, we shall later consider processes with a continuous time parameter. The state space of our processes can be either discrete or continuous and either real-valued or vector-valued. The index  $n$  for the sequence is meant to correspond to the physical parameter mentioned above. In some models, however, the sequence index will not be associated with the positive integers. With this set-up given, our aim is to scale the time parameter and to translate and scale the space variable in such a

- 
1. This work was supported by the Office of Naval Research, Contract Nonr-401(55).
  2. This paper was prepared for the American Mathematical Society 1967 Summer Seminar on Mathematics of the Decision Sciences to be held at Stanford University from July 10 to August 11, 1967.

manner that the resulting processes converge to a limit process as  $n$  goes to infinity. In general we shall seek sequences  $\{a_n\}$ ,  $\{b_n\}$ , and  $\{c_n\}$  to form the processes<sup>3</sup>

$$Y_n(t) = \frac{X_n([a_n t]) - b_n}{c_n}, \quad t \geq 0, \quad n = 1, 2, \dots$$

The  $a_n$ 's and  $c_n$ 's will be positive real numbers, generally tending to infinity as  $n \rightarrow \infty$ . However, the  $b_n$ 's will be vectors if the  $X_n$ 's are. Notice that the  $Y_n(\cdot)$  processes will be constant for stretches of length  $a_n^{-1}$  because of the discrete nature of the  $X_n$ 's and hence have discontinuous paths. There is an alternative way to define the  $Y_n(\cdot)$  processes which leads to continuous path functions. This latter approach has certain advantages and will be introduced later.

There are a number of modes of convergence for the  $Y_n(\cdot)$  processes which are of interest. The simplest of these is to require convergence in distribution of the value of the  $Y_n(\cdot)$  processes at a fixed value of  $t$ ; i.e.,

$$\lim_{n \rightarrow \infty} P_n\{Y_n(t) \leq x\} = P\{Y(t) \leq x\} \quad \text{for all } t \geq 0$$

and all  $x$  for which the right-hand side is continuous, where  $Y(\cdot)$  is a limit process defined on a probability space  $(\Omega, \mathcal{F}, P)$ . Often the limit process,  $Y(\cdot)$ , is a diffusion; i.e., a strong Markov process with continuous path functions. The classical central limit theorem is of this type, where  $Y(\cdot)$  is Brownian motion. Next we might be interested in showing the convergence of the finite-dimensional distributions (f.d.d.) of the  $Y_n(\cdot)$  processes to the corresponding distributions of the  $Y(\cdot)$  process. In other words, for every  $k > 1$

---

3. The symbol  $[x]$  denotes the integer part of  $x$ .

and  $0 \leq t_1 < t_2 < \dots < t_k$  we would like to show that the

$$\lim_{n \rightarrow \infty} P_n \{Y_n(t_1) \leq x_1, \dots, Y_n(t_k) \leq x_k\} = P\{Y(t_1) \leq x_1, \dots, Y(t_k) \leq x_k\}.$$

for all  $x_j$  for which the right-hand side is continuous.

A third mode of convergence which is important for applied problems is weak convergence of a sequence of probability measures defined as follows. Let  $S$  be a metric space and  $\mathcal{B}$  be the smallest Borel field containing the open sets of  $S$ . If  $\nu_n$  and  $\nu$  are probability measures on  $\mathcal{B}$  and if the

$$\lim_{n \rightarrow \infty} \int_S f d\nu_n = \int_S f d\nu$$

for every bounded, continuous, real-valued function  $f$  on  $S$ , then we say  $\nu_n$  converges weakly to  $\nu$  and write  $\nu_n \Rightarrow \nu$ .

For our problems the measures  $\nu_n$  will be generated by the  $Y_n(\cdot)$  processes and the measure  $\nu$  by  $Y(\cdot)$ . In most cases the metric space  $S$  will be  $C[0,1]$ , the space of continuous functions on  $[0,1]$  with the metric of uniform convergence, or a multi-dimensional analog. Weak convergence is important because it implies convergence in distribution of certain functionals of the original processes. Sometimes in applications the distribution of the functional is more important than that of the original process.

Once these limit theorems have been obtained we can consider using the limit distributions as an approximation to the distribution of  $X_n(\cdot)$  when the parameter  $n$  is sufficiently "large." Of course, the question of how large  $n$  must be before a "good" approximation is obtained introduces the question of rates of convergence for the limit theorems. Little work of an analytical nature has been done on the rate of convergence issue, however, for some models numerical results are available.

The prototype of these limit theorems and their resulting approximations is the convergence of sums of independent, identically distributed (i.i.d.) random variables to Brownian motion. These results will be sketched in Section 2. Models from queueing theory will be taken up in Sections 3, 4, and 5. Section 6 will be concerned with the multi-urn Ehrenfest model. A quality control model will be discussed in Section 7. Finally, in Section 8 brief mention will be made of other work on the convergence of processes in applied probability.

## 2. Sums of Random Variables and Brownian Motion

Let  $X_1, X_2, \dots$  be a sequence of i.i.d. random variables defined on the product probability space  $(\Omega, \mathcal{F}, P)$  and having mean 0 and variance 1. We shall denote the partial sums by  $S_i = X_1 + \dots + X_i$  and set  $S_0 = 0$ . The appropriate sequence of processes to consider is defined by

$$Y_n(t) = \frac{S_{[nt]}}{\sqrt{n}}, \quad 0 \leq t \leq 1, \quad n = 1, 2, \dots$$

The fact that we have restricted the time parameter  $t$  to the unit interval is not important, but does make the exposition easier. Observe that for a fixed value of  $t$  the number of jumps of the  $Y_n(\cdot)$  process is of order  $n$ , the mean value of each jump is 0, and the variance of each jump is  $n^{-1}$ . Thus the jumps are occurring very often and their heights are becoming small as  $n$  grows large. Hence it is natural to hope that as  $n \rightarrow \infty$  the paths of  $Y_n(\cdot)$  will become continuous.

Alternatively, we could consider a sequence of processes defined by

$$Z_n(t) = \frac{S_k}{\sqrt{n}} + (nt-k) \frac{X_{k+1}}{n}, \quad kn^{-1} \leq t \leq (k+1)n^{-1}$$

for  $k = 0, 1, \dots, n-1$ . Notice that  $Z_n(t) = Y_n(t)$  for  $t$  of the form  $kn^{-1}$ ,  $k = 0, \dots, n$ , and is defined by linear interpolation for other values of  $t$ . Hence the paths of  $Z_n(\cdot)$  are automatically continuous and this is convenient when considering weak convergence of probability measures.

We shall be interested in showing that the sequences  $\{Y_n(\cdot)\}$  and  $\{Z_n(\cdot)\}$  converge, as discussed in Section 1, to Brownian motion,  $Y(\cdot)$ , on the unit interval. As a check to see that we are on the right track, we calculate the infinitesimal mean and variance of  $Y_n(t)$  per unit time. Denoting these means and variances by  $m_n(y)$  and  $\sigma_n^2(y)$ , we have

$$m_n(y) = n E\{Y_n(t + \frac{1}{n}) - Y_n(t) | Y_n(t) = y\}$$

$$= n E \frac{X_{[nt]+1}}{\sqrt{n}} = 0$$

and

$$\begin{aligned}\sigma_n^2(y) &= n E\{[Y_n(t+\frac{1}{n}) - Y_n(t)]^2 | Y_n(t) = y\} \\ &= E\{X_{[nt]+1}^2\} = 1.\end{aligned}$$

This infinitesimal mean and variance agrees with those of Brownian motion which provides a useful check before proceeding to a rigorous analysis.

The proof of the convergence of the one-dimension distributions is simply the central limit theorem; i.e.,

$$\lim_{n \rightarrow \infty} P\{Y_n(t) \leq x\} = (2\pi t)^{-1/2} \int_{-\infty}^x \exp\{-\frac{s^2}{2t}\} ds.$$

The normal distribution on the right-hand side is also the distribution of the position of Brownian motion at time  $t$ . To show convergence of the f.d.d. the Lévy continuity theorem for characteristic functions can be used. The case  $k = 2$  embodies the general argument which goes as follows. Let  $\phi_n(s_1, s_2; t_1, t_2)$  be the joint characteristic function of  $Y_n(t_1)$  and  $Y_n(t_2)$ . Then

$$\begin{aligned}\phi_n(s_1, s_2; t_1, t_2) &= E\left[\exp\left(\frac{is_1 + s_2}{\sqrt{n}} S_{[nt_1]} + \frac{is_2}{\sqrt{n}} (S_{[nt_2]} - S_{[nt_1]})\right)\right] \\ &= E\left[\exp\left(\frac{is_1 + s_2}{\sqrt{n}} S_{[nt_1]}\right)\right] \cdot E\left[\exp\left(\frac{is_2}{\sqrt{n}} S_{[nt_2] - [nt_1]}\right)\right].\end{aligned}$$

Letting  $n \rightarrow \infty$  we see that

$$\phi_n(s_1, s_2; t_1, t_2) \rightarrow \exp\left\{-\frac{(s_1 + s_2)^2 t_1}{2} - \frac{s_2^2 (t_2 - t_1)}{2}\right\},$$

which is the joint characteristic function of Brownian motion at times  $t_1$  and  $t_2$ . Similar arguments can be used to show the convergence of the f.d.d. of the  $Z_n(\cdot)$  processes.

To establish weak convergence of the measures associated with  $\{Y_n(\cdot)\}$  or  $\{Z_n(\cdot)\}$  two steps are required. The first step is the convergence of the f.d.d. which we have demonstrated above. Secondly, the probability that the approximating processes can have large fluctuation between points at which they are determined by their f.d.d. must be shown to be small. The notation of weak convergence is intimately related to the so-called invariance principles. An invariance principle for sums of i.i.d. random variables states roughly that the limit of the distribution of various functionals of the  $S_i$ 's is independent of the common distribution of the  $X_i$ 's, provided the mean is 0 and variance 1. Such an invariance principle was first given by Erdos and Kac (1946) and later generalized by Donsker (1951) and Billingsley (1956). To carry out the second step mentioned above for  $\{Y_n(\cdot)\}$  (respectively,  $\{Z_n(\cdot)\}$ ) the reader should consult Donsker (1951) (Billingsley (1956)). Other important references for weak convergence of sums of i.i.d. random variables are Prokhorov (1956), Skorokhod (1956), and Itô-McKean (1965).

It is important to note that while the approximating processes could all be defined on a single probability space the limiting Brownian motion was defined on another probability space. Thus with this set-up it makes no sense to speak of convergence with probability one or in quadratic mean, for example. However, when the  $X_i$ 's are Bernoulli random variables, Knight (1962) has succeeded in defining the approximating processes and the limit process on a single probability space and then shown probability one convergence.

### 3. Waiting Time for the Queue GI/G/1

In this section we shall consider the distribution of the waiting time in the single-server queue with general independent input and general service time. Our objective is to demonstrate the usefulness of the notion of weak convergence in studying the asymptotic behavior of the waiting time as the traffic intensity  $\rho$ , goes to 1. This discussion is based on work of Prokhorov (1956, 1963).

Let  $W_n$  be the waiting time (time before service begins) for the  $n^{\text{th}}$  customer. We shall denote the service time of the  $n^{\text{th}}$  customer by  $v_n$  and the interval between the arrival times of the  $n^{\text{th}}$  and  $(n+1)^{\text{st}}$  customers by  $u_n$ . Then if  $W_1 = 0$ ,

$$(1) \quad F_{n+1}(x) = \Pr\{W_{n+1} \leq x\} = \Pr\{S_1 \leq x, S_2 \leq x, \dots, S_n \leq x\},$$

where  $X_n = v_n - u_n$  and  $S_i = \sum_{j=1}^i X_j$ . This result was first derived by Lindley

(1952); see Prabhu (1965) for a comprehensive discussion of the GI/G/1 queue.

The usual independence assumptions regarding  $\{v_n\}$  and  $\{u_n\}$  imply that the  $X_i$ 's are independent. Since the distribution of  $W_{n+1}$  is seen to be equal to the distribution of the maximum functional on the process of partial sums,  $S_i$ , it is natural to hope for limit theorems which lead to the maximum functional on Brownian motion.

Consider now a sequence of such queueing processes indexed by  $\delta_i$  in which  $E(u_n^{(\delta_i)}) = 1/(1-\delta_i)$ ,  $E(v_n^{(\delta_i)}) = 1$ , and hence  $\rho_i = 1-\delta_i$ . Thus

$$E(X_n^{(\delta_i)}) = -\delta_i / (1-\delta_i). \quad \text{We shall treat the case } \delta_i > 0 \text{ (or } \rho_i < 1) \text{ here,}$$

although for many of the arguments the sign of  $\delta_i$  is not important. Our goal

now is to obtain limit theorems for  $W_{n_i}^{(\delta_i)}$  as  $\delta_i \searrow 0$  and  $n_i \rightarrow \infty$ . There are a number of possible limit theorems depending on whether  $n_i \delta_i^2$  converges to zero, a positive constant, or plus infinity.

Theorem 3.1 of Prokhorov (1956) is the principal tool used in obtaining these limit theorems. The set-up for the theorem is as follows. A double sequence

$$X_{n,1}, X_{n,2}, \dots, X_{n,k_n}$$

of random variables is given which are independent for each  $n$  and subject to the asymptotic negligibility condition

$$\lim_{n \rightarrow \infty} \max_{1 \leq k \leq k_n} \Pr\{|X_{n,k}| > \epsilon\} = 0 \quad \text{for all } \epsilon > 0.$$

The first two moments satisfy  $E\{X_{n,k}\} = 0$ ,  $\sigma_{n,k}^2 = \sigma^2[X_{n,k}] > 0$ , and  $\sum_{k=1}^{k_n} \sigma_{n,k}^2 = 1$ .

Let the partial sums be denoted by  $S_{n,0} = 0$  and  $S_{n,k} = \sum_{j=1}^k X_{n,j}$  for  $1 \leq k \leq k_n$ . We let  $t_{n,k} = \sigma^2[S_{n,k}]$  and construct the continuous path function  $Y_n(t)$  which is piece-wise linear with vertices at the points  $(t_{n,k}, S_{n,k})$ . The measure induced by  $Y_n(t)$  on  $C[0,1]$  we shall denote by  $P_n$  and that induced by Brownian motion by  $P$ . Then the theorem is as follows.

A necessary and sufficient condition for  $P_n$  to converge weakly to  $P$  is that

$$(2) \quad \lim_{n \rightarrow \infty} \sum_{k=1}^{k_n} \int_{|x| > \lambda} x^2 dF_{n,k}(x) = 0$$

for all  $\lambda > 0$ , where  $F_{n,k}(x) = \Pr\{X_{n,k} \leq x\}$ .

First consider the case where  $n_i \delta_i^2 \rightarrow 0$  as  $i \rightarrow \infty$ . We shall assume that  $\sigma^2[X_j^{(\delta_i)}] = \sigma_i > 0$ ,  $\sigma_i \rightarrow \sigma$ , and that

$$(3) \quad \int_{|x| \geq z} x^2 dK^{(\delta_i)}(x) \rightarrow 0$$

uniformly in  $\delta_i$  as  $z \rightarrow \infty$ , where  $K^{(\delta_i)}$  is the distribution function of  $X_n^{(\delta_i)}$ .

The latter condition assures us that condition (2) is satisfied for our queueing problem. From (1) we have for  $0 \leq x < \infty$

$$(4) \quad F_{n_i+1}^{(\delta_i)}(x\sigma_i\sqrt{n_i}) = \Pr\{W_{n_i+1}^{(\delta_i)} \leq x\sigma_i\sqrt{n_i}\} = \Pr\left\{\frac{S_k^{(\delta_i)}}{\sigma_i\sqrt{n_i}} \leq x, 1 \leq k \leq n_i\right\}.$$

One can easily check to see that the conditions of the above theorem hold and hence

$$(5) \quad \Pr \left\{ \frac{S_k^{(\delta_i)}}{\sigma_i \sqrt{n_i}} \leq x + \frac{k E[X_1^{(\delta_i)}]}{\sigma_i \sqrt{n_i}}, \quad 1 \leq k \leq n_i \right\}$$

converges as  $i \rightarrow \infty$  to the maximum functional of Brownian motion, namely

$$P \left( \max_{0 \leq t \leq 1} Y(t) \leq x \right),$$

where  $Y(t)$  is the path function of Brownian motion. This probability is known to be  $2\Phi(x) - 1$ , the truncated normal distribution. Since we have assumed that  $n_i \delta_i^2 \rightarrow 0$ , the terms

$$(6) \quad \frac{k E[X_1^{(\delta_i)}]}{\sigma_i \sqrt{n_i}} = - \frac{k \delta_i}{\sigma_i \sqrt{n_i} (1 - \delta_i)} \rightarrow 0 \quad \text{as } i \rightarrow \infty,$$

for  $k = 1, 2, \dots, n_i$ . Hence putting together (5) and (6) yields

$$F_{n_i+1}^{(\delta_i)}(x \sigma_i \sqrt{n_i}) \rightarrow 2\Phi(x) - 1.$$

This result is of course also true in the case where  $\delta_i = 0$ .

Next we assume  $n_i \delta_i^2 \rightarrow \tau$ ,  $0 < \tau < \infty$ . Using (1) again we have

$$F_{n_i+1}^{(\delta_i)}(x/\delta_i) = \Pr \left\{ \frac{S_k^{(\delta_i)} - k E[X_1^{(\delta_i)}]}{\sigma_i \sqrt{n_i}} \leq \frac{x - \frac{k}{n_i} \cdot n_i E[X_1^{(\delta_i)}] \delta_i}{\sigma_i \sqrt{n_i} \delta_i}, \quad 1 \leq k \leq n_i \right\}.$$

By assumption  $-n_i E[X_1^{(\delta_i)}] \delta_i \rightarrow \tau$  and  $\sigma_i \sqrt{n_i} \delta_i \rightarrow \sigma \sqrt{\tau}$ . Using the weak convergence again we have

$$(7) \quad F_{n_i+1}^{(\delta_i)}(x/\delta_i) \rightarrow P \left( \max_{0 \leq t \leq 1} [Y(t) - \frac{t\tau}{\sigma\sqrt{\tau}}] \leq \frac{x}{\sigma\sqrt{\tau}} \right).$$

By a change in time scale the last probability can also be written as

$$(8) \quad P\left(\max_{0 \leq t \leq \tau} [Y(t) - \frac{t}{\sigma}] \leq \frac{x}{\sigma}\right) .$$

Finally, consider the case where  $n_i \delta_i^2 \rightarrow \infty$ . We would expect the limit of  $F_{n_i+1}^{(\delta_i)}(x/\delta_i)$  to be the expression in (8) with  $\tau$  replaced by  $\infty$ . This is in fact true, although an additional argument employing Kolmogorov's inequality is required. For a fixed  $\delta > 0$ , the stationary distribution of the Markov chain  $\{W_n^{(\delta)}\}$  exists and is given by

$$F^{(\delta)}(x) = P\left(\sup_{k \geq 1} S_k^{(\delta)} \leq x\right) .$$

In the course of deriving (8) with  $\tau$  replaced by  $\infty$ , Prokhorov also shows that

$$(9) \quad F_{n_i+1}^{(\delta_i)}(x/\delta_i) \rightarrow P\left(\max_{0 \leq t < \infty} [Y(t) - t\sigma^{-1}] \leq x\sigma^{-1}\right) = 1 - e^{-2x/\sigma^2} .$$

The last equality was derived by Darling and Siegert (1953). The result given in (9) was first derived by Kingman (1962). For an expository account of Kingman's work in this area of so-called "heavy traffic" the reader should consult Kingman (1965).

From this work of Prokhorov's it should be clear that the notion of weak convergence is an important one for applied probability. For other work in this spirit consult Viskov (1964), Viskov and Prokhorov (1964), and Borovkov (1964).

#### 4. The Many Server Queue and Telephone Trunking Problem

In this section we shall discuss the many server queue and telephone trunking problem (infinite server queue) with Poisson arrivals and exponential service time. As usual the service times are independent of the arrival process and no server is idle if a customer is waiting. If there are  $n$  servers and we let  $X_n(t)$  denote the number of customers waiting or being served at time  $t$ , then it is well known that  $X_n(t)$  is a birth and death process. As such

the transition probabilities,  $p_{ij}(t) = P\{X_n(t+\tau) = j | X_n(\tau) = i\}$ , can be expressed by the integral representation of Karlin and McGregor (1958). If we knew  $p_{ij}(t)$  explicitly, we would in principle be able to calculate all the distributions of interest for this model. Unfortunately, it is exceedingly difficult to compute  $p_{ij}(t)$  when  $n$  is larger than 3 or 4.

Motivated by this difficulty, we shall seek a diffusion approximation for  $X_n(t)$  when  $n$  is large. We shall follow the discussion of the author (Iglehart (1965)). Naturally, if we keep the expected interarrival time and the expected service time fixed, the behavior of the many server queue approaches that of the telephone trunking problem. While the transition probabilities for the telephone trunking problem are well known, we would have lost the characteristics of the original problem. In fact, the traffic intensity  $\rho$ , defined to be the ratio of the arrival rate to  $n$  times the service rate, would tend to zero whereas the processes being approximated have a positive traffic intensity. Therefore, we shall let the arrival rate become large with the number of servers, and keep the service time constant in such a manner that  $\rho$  is maintained at a fixed value less than one. To be specific we let  $X_n(t)$  be the birth and death process with parameters

$$\begin{aligned} \lambda_i^{(n)} &= n\rho \\ \mu_i^{(n)} &= \begin{cases} i & i \leq n \\ n & n > i \end{cases} \end{aligned}$$

for  $i = 0, 1, \dots$  and  $n = 1, 2, \dots$ , where  $0 < \rho < 1$ .

As a heuristic aid to setting up the appropriate approximating processes, consider the imbedded jump process of  $X_n(t)$ . The expected displacement in one jump starting at state  $i \leq n$  is  $(n\rho - i)/(n\rho + i)$  which is non-negative if  $i \leq n\rho$  and negative if  $i > n\rho$ . For  $i > n$  the expected displacement is always

negative. In other words, the state  $[n\rho]$  is an "equilibrium point" of the process. Thus for our approximating processes it is natural to consider the fluctuations of  $X_n(t)$  about  $[n\rho]$ , measured in an appropriate scale. We shall set

$$Y_n(t) = \frac{X_n(t) - n\rho}{(n\rho)^{1/2}}, \quad n = 1, 2, \dots$$

Again as a guide to the limiting diffusion process, we shall calculate the infinitesimal mean and variance of  $Y_n(t)$ . The infinitesimal mean is defined in terms of the jump process  $\tilde{Y}_n(k)$ , say, as

$$m_n(y) = E\{\tilde{Y}_n(k+1) - \tilde{Y}_n(k) \mid \tilde{Y}_n(k) = y\} / E\{\text{holding time in } y\}.$$

In our case if we let  $\alpha_n(y) = [n\rho + (n\rho)^{1/2}y]$ , then

$$m_n(y) = \frac{1}{(n\rho)^{1/2}} \frac{\lambda_{\alpha_n(y)}^{(n)} - \mu_{\alpha_n(y)}^{(n)}}{\lambda_{\alpha_n(y)}^{(n)} + \mu_{\alpha_n(y)}^{(n)}} \cdot (\lambda_{\alpha_n(y)}^{(n)} + \mu_{\alpha_n(y)}^{(n)}).$$

We now take  $n$  so large that  $\alpha_n(y) < n$  (which is possible since  $\rho < 1$ ). Thus

$$(10) \quad m_n(y) = \frac{1}{(n\rho)^{1/2}} \{n\rho - [n\rho + (n\rho)^{1/2}y]\} \rightarrow -y \quad \text{as } n \rightarrow \infty.$$

The infinitesimal variance is similarly defined as

$$(11) \quad \sigma_n^2(y) = E\{(\tilde{Y}_n(k+1) - \tilde{Y}_n(k))^2 \mid \tilde{Y}_n(k) = y\} / E\{\text{holding time in } y\}.$$

Hence

$$\begin{aligned} \sigma_n^2(y) &= \frac{1}{(n\rho)} \frac{\lambda_{\alpha_n(y)}^{(n)} + \mu_{\alpha_n(y)}^{(n)}}{\lambda_{\alpha_n(y)}^{(n)} + \mu_{\alpha_n(y)}^{(n)}} \cdot (\lambda_{\alpha_n(y)}^{(n)} + \mu_{\alpha_n(y)}^{(n)}) \\ &= \frac{1}{n\rho} (n\rho + [n\rho + (n\rho)^{1/2}y]) \rightarrow 2 \quad \text{as } n \rightarrow \infty. \end{aligned}$$

The limit process should then be governed by the backward equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} - x \frac{\partial u}{\partial x}, \quad -\infty < x < \infty,$$

which is recognized as the equation of the Ornstein-Uhlenbeck diffusion process.

At this point we are in need of a technique for showing that the distribution of  $Y_n(t)$  converges to the distribution of the Ornstein-Uhlenbeck process  $Y(t)$ , say. If we could explicitly obtain the representation for  $p_{ij}(t)$ , we might use analytic techniques involving orthogonal polynomials to achieve our result. While the parameters  $\{\pi_j\}$  and the orthogonal polynomials  $\{Q_i(x)\}$  can be easily obtained, the measure  $\psi(x)$  is difficult to characterize. In fact, the difficulty in obtaining this representation provided our initial motivation for looking for a diffusion approximation. Fortunately, there is a general theory due to Stone (1961, 1963) which gives necessary and sufficient conditions for the weak convergence of a sequence of birth and death processes (or random walks, or diffusions) to a limiting diffusion process. Although the general set-up and notation required to discuss Stone's results in detail are too involved for this paper, perhaps a few heuristic remarks would be helpful.

The processes considered by Stone (random walks, birth and death processes, and diffusions) enjoy the property that points in the state space are not jumped over by the process; i.e., possible transitions can only occur to neighboring states in the discrete case and the path functions are continuous in the continuous state space case. This property results in the infinitesimal operator of the semi-group of transition functions being a local operator. The infinitesimal operator is essentially determined (aside from any boundary conditions which may have to be imposed) by the infinitesimal mean and variance of the process. Since convergence of the infinitesimal operators of a sequence of processes implies convergence of the processes, it seems natural to assume that convergence of the

infinitesimal means and variances would also imply convergence of the processes. This is in fact true, except the infinitesimal operator of a semi-group is not determined until the boundary conditions (lateral conditional in Feller's (1957) terminology) have been specified. Hence to obtain convergence of the processes we need to assume that the behavior of the sequence of processes in the neighborhood of a boundary point convergence to that of the limit process. Finally, we need to check that the state space of the approximating processes becomes dense in the state space of the limit process.

In our example of the many server queue we have shown in (10) and (11) that the infinitesimal mean and variance convergence uniformly in every compact interval of the line. Furthermore, the state space of  $Y_n(t)$  becomes dense in  $(-\infty, +\infty)$ , the state space of the limit process, as  $n \rightarrow \infty$ . These are the essential facts required to apply Stone's results (cf. Iglehart (1965)). The conclusion is that the processes  $Y_n(t)$  converge weakly to  $Y(t)$  provided  $X_n(0) = [n\rho + (n\rho)^{1/2}y]$  for any real  $y$ .

To illustrate how the representation for  $p_{ij}(t)$  can be used to obtain a limiting diffusion, we shall consider the telephone trunking problem. In this model  $X_n(t)$ , the number of busy channels, is a birth and death process with parameters

$$\begin{aligned}\lambda_j^{(n)} &= nc \\ \mu_j^{(n)} &= j\end{aligned}$$

for  $j = 0, 1, \dots$ , and  $n = 1, 2, \dots$ , where  $c > 0$ . The representation (Karlin and McGregor (1958)) in this case is

$$p_{ij}(t) = \pi_j \sum_{k=0}^{\infty} e^{-kt} c_i(k; nx) c_j(k; nc) \frac{e^{-nc} (nc)^k}{k!},$$

where  $\{\omega_k(x; a)\}$  are the Poisson-Charlier polynomials, orthogonal with respect

to the measure  $e^{-a} a^x / x!$  on  $x = 0, 1, \dots$  and  $\pi_j = (nc)^j / j!$ . By showing that the Poisson-Charlier polynomials, properly normalized, converge to the Hermite polynomials (this is not surprising considering the fact that the Hermite polynomials are orthogonal with respect to the normal density and the implications of the central limit theorem), it is not hard to show that for

$$\alpha_n(x) = [(nc)^{1/2} x + nc] p_{\alpha_n(x), \alpha_n(y)}(t) \text{ is asymptotic (when normalized)}$$

to the transition density for the Ornstein-Uhlenbeck process. This convergence provides a local limit theorem for the convergence of the processes

$$(X_n(t) - nc) / (nc)^{1/2} \text{ to the Ornstein-Uhlenbeck process.}$$

For another example of such a local limit theorem obtained from the integral representation for  $p_{ij}(t)$  in the case of the Ehrenfest urn model, the reader should consult Karlin and McGregor (1965).

### 5. Luchak's Queueing Model in Heavy Traffic

Luchak (1958) studied a single-server queueing model in which customers arrive according to a Poisson process with rate  $\lambda > 0$  and require a random number,  $N$ , of phases of service, each phase being exponential with parameter  $\mu$ . Luchak considered the transient behavior of the phase length process,  $Q(t)$ , which is the number of phases present in the system at time  $t$  including the phase in service. His result is given in terms of a rather unwieldy transform.

Recently Lalchandani (1967) considered in his thesis the behavior of  $Q(t)$  in heavy traffic (i.e., as the traffic intensity approaches one). We shall give a brief outline of his method which seems to have some general interest.

The distribution of  $N$  is discrete ( $P[N=n] = c_n, n=1,2,\dots$ ) with finite mean,  $a$ , and second moment,  $b$ . With these parameters the traffic intensity  $\rho = \lambda a / \mu$ . Consider now a sequence of queueing systems indexed by  $n$  ( $n \geq 1$ ) for which the arrival and service parameters are  $\lambda_n$  and  $\mu_n$ . For the corresponding traffic intensity,  $\rho_n$ , to tend to 1 we choose, as an example,  $\lambda_n = n$  and

$\mu_n = a(n + \sqrt{n})$ . If  $Q_n(t)$  denotes the phase length process for the  $n^{\text{th}}$  system, Lalchandani shows that the distribution of  $X_n(t)$ , defined as

$$X_n(t) = \frac{Q_n(t) - n}{[(a+b)n]^{1/2}},$$

converges as  $n \rightarrow \infty$  to the distribution of Brownian motion at time  $t$  with initial state  $y$ , provided  $Q_n(0) = [\sqrt{(a+b)n} \cdot y + n]$ .

While the  $Q_n(t)$  process is a continuous parameter Markov chain, it is not a birth and death process and hence we do not have available the special techniques for such processes. Consider, however, the discrete parameter jump chain  $\tilde{Q}_n(k)$ , corresponding to  $Q_n(t)$ . In one step the chain  $\tilde{Q}_n(k)$  goes up  $j$  with probability  $c_j \lambda_n / (\lambda_n + \mu_n)$  and down 1 with probability  $\mu_n / (\lambda_n + \mu_n)$ , provided  $\tilde{Q}_n(k) > 0$ . If  $\tilde{Q}_n(k) = 0$ , then the only transitions are up  $j$  with probability  $c_j$ . Except for the troublesome state 0, the  $\tilde{Q}_n(k)$  process is a process of sums of independent random variables. However, when  $\rho_n \nearrow 1$ , we would not expect the queue to be idle often, and thus it is not too surprising that the limit process for  $\tilde{X}_n([\lambda_n + \mu_n]t)$  is Brownian motion, where

$\tilde{X}_n(k) = (\tilde{Q}_n(k) - n) / [(a+b)n]^{1/2}$ . One of the crucial steps in showing this convergence in distribution is the study of the behavior of  $P\{\tilde{Q}_n(j) = 0\}$ .

Once the convergence of the suitably translated and scaled jump chain is obtained it is not difficult to show that the  $X_n(t)$  process converges. If we consider the time points at which jumps in the  $Q_n(t)$  process occur, it is clear that these time points are essentially a renewal process with exponential lifetimes having parameter  $(\lambda_n + \mu_n)$ . Very occasionally, however, the queue is empty and the exponential parameter is  $\lambda_n$ . Define  $N_n(t)$  to be the number of jumps of the  $Q_n(\cdot)$  process in the interval  $[0, t]$ . Then the mean and variance of  $N_n(t)$  are shown to behave like  $(\lambda_n + \mu_n)t$  as  $n \rightarrow \infty$ . Now a conditional probability argument can be used in which the  $P\{X_n(t) \leq x | N_n(t) = j\}$  is

is identified with the  $P(\tilde{X}_n(j) \leq x)$ . Finally, in this manner Lalchandani shows that the distribution of  $X_n(t)$  converges to the appropriate normal distribution.

While the argument outlined above used the specific structure of the Markov chain  $Q_n(t)$ , it seems likely that the idea of looking first for limit processes of the jump chain should have greater applicability.

#### 6. The Multi-urn Ehrenfest Model

In the multi-urn Ehrenfest model  $N$  balls are distributed among  $d+1$  ( $d \geq 2$ ) urns. If we label the urns  $0, 1, \dots, d$ , then the system is said to be in state  $\mathbf{i} = (i_1, i_2, \dots, i_d)$  when there are  $i_j$  balls in urn  $j$  ( $j=1, 2, \dots, d$ ) and  $N - \mathbf{i} \cdot \mathbf{1}$  balls in urn 0. At discrete epochs a ball is chosen at random from one of the  $d+1$  urns; each of the  $N$  balls has probability  $1/N$  of being selected. The ball chosen is removed from its urn and placed in urn  $i$  ( $i=0, 1, \dots, d$ ) with probability  $p^i$ , where the  $p^i$ 's are elements of a given vector,  $(p^0, p)$ , satisfying  $p^i > 0$  and  $\sum_{i=0}^d p^i = 1$ . We shall let  $X_N(k)$  denote the state of the system after the  $k^{\text{th}}$  such rearrangement of balls. In this section we shall discuss some limit theorems, which were obtained by the author (Iglehart (1967)) for the sequence of processes  $\{X_N(k): k=0, \dots, N\}$  as  $N$  tends to infinity. For the classical Ehrenfest model ( $d=1, p^0=p^1=1/2$ ) Kac (1947) showed that the distribution of  $(X_N([Nt]) - N/2)/(N/2)^{1/2}$  converges as  $N \rightarrow \infty$  to the distribution of the Ornstein-Uhlenbeck process at time  $t$  having started at  $y_0$  at  $t=0$ , provided  $X_N(0) = [(N/2)^{1/2}y_0 + N/2]$ . Recently, Karlin and McGregor (1965) obtained a similar result for the continuous time version of the model with  $d=2$ ; in this version the random selection of balls is done at the occurrence of events of an independent Poisson process.

---

4. The vector  $\mathbf{1}$  has all its components equal to 1 and  $\mathbf{x} \cdot \mathbf{x}$  is the usual scalar product.

A preliminary calculation indicates that the process  $\{\chi_N(k): k=0, \dots, N\}$  is attracted to the pseudo-equilibrium state  $Np$  and that states far from  $Np$  will only occur rarely. Thus it is natural to consider the fluctuations of  $\chi_N(k)$  about  $Np$  measured in an appropriate scale. For our purposes the appropriate processes to consider are  $\{\chi_N(k): k=0, \dots, N\}$ , where

$$\chi_N(k) = (\chi_N(k) - Np)/N^{1/2}.$$

Next we define a sequence of stochastic processes  $\{\chi_N(t): 0 \leq t \leq 1\}$  which are continuous, linear on the intervals  $((k-1)N^{-1}, kN^{-1})$ , and satisfy  $\chi_N(kN^{-1}) = \chi_N(k)$  for  $k=0, 1, \dots, N$ . In other words we let

$$\chi_N(t) = \chi_N(k) + (Nt-k)(\chi_N(k+1) - \chi_N(k))$$

if  $kN^{-1} \leq t \leq (k+1)N^{-1}$ . Throughout this discussion we shall let<sup>5</sup>

$\chi_N^i(0) = [N^{1/2}y_0^i + Np^i]$ , where  $\chi_0 = (y_0^1, \dots, y_0^d)$  is an arbitrary, but fixed, element of<sup>6</sup>  $R^d$ . With this initial condition and the Markov structure of the model, the processes  $\{\chi_N(k): k=0, \dots, N\}$  for  $N=1, 2, \dots$  can be defined on a probability triple  $(\Omega_N, \mathcal{F}_N, P_N)$ . We shall let  $C_d[0,1]$  denote the product space of  $d$  copies of  $C[0,1]$ , the space of continuous functions on  $[0,1]$  with the topology of uniform convergence, and endow  $C_d[0,1]$  with the product topology. The topological Borel field of  $C_d[0,1]$  will be denoted by  $\mathcal{C}_d$ . Clearly, the transformation taking the sequence  $\{\chi_N(k): k=0, \dots, N\}$  into  $\{\chi_N(t): 0 \leq t \leq 1\}$  is measurable and induces a probability measure on  $\mathcal{C}_d$ . We shall denote this induced measure by  $\mu_N(\cdot; \chi_0)$ .

5. It will always be understood that  $N$  is sufficiently large so that

$0 \leq \chi_N^i(0) \leq N$  for all  $i=1, 2, \dots, d$ , where  $\chi_N^i(\cdot)$  is the  $i$ -th component of the vector  $\chi_N(\cdot)$ .

6.  $R^d$  is  $d$ -dimensional Euclidean space.

The principal result of Iglehart (1967) is that  $\mu_N(\cdot; y_0) \Rightarrow \mu(\cdot; y_0)$  as  $N \rightarrow \infty$ , where  $\mu(\cdot; y_0)$  is the probability measure on  $C_d$  of a d-dimensional diffusion process,  $y(\cdot)$ , starting at the point  $y_0$ . The limit process  $y(\cdot)$  is a d-dimensional analog of the Ornstein-Uhlenbeck process whose distribution at time  $t$  is a multi-variate normal with mean vector  $e^{-t}y_0$  and covariance matrix  $\Sigma_t$ , where the elements of  $\Sigma_t$  are

$$\sigma_{ij} = \begin{cases} (1-e^{-2t})p^i(1-p^i) & , \quad i=j \\ -(1-e^{-2t})p^ip^j & , \quad i \neq j. \end{cases}$$

To obtain the weak convergence of the measures  $\mu_N$  to  $\mu$  we must first show the convergence of the corresponding f.d.d. We shall only be able to sketch the proof here. For the convergence of the distribution of  $y_N(t)$  we can consider the distribution of  $Y_N([Nt])$ , since  $|Y_N^i([Nt]) - y^i(t)| \leq N^{-1/2}$  for  $i=1, \dots, d$  with probability one. The method of characteristic functions and the Lévy continuity theorem is used. If we let  $\psi_N(z; k) = E_N(\exp(iz \cdot Y_N(k)))$ , then by using a standard conditional probability argument and obvious asymptotic expansions we show that

$$\psi_N(z, k+1) = E_N(z) \psi_N(h_N(z, 1), k)$$

for  $k=0, 1, \dots, N-1$ , where

$$E_N(z) = \exp(-N^{-1}z'Az + o(N^{-1}))$$

$$A = (1-e^{-2t})^{-1} \Sigma_t$$

$h_N(z, 1) = (1-N^{-1} + o(N^{-1}))z$  as  $N \rightarrow \infty$ , and the terms  $o(N^{-1})$  are uniform for  $z$  in a compact set of  $R^d$  and independent of  $k$ . From this result, a simple iteration one shows that

7. The symbol  $E_N(\cdot)$  denotes expectation with respect to  $P_N$ .

$$(12) \quad \psi_N(x, k) = \prod_{j=0}^{k-1} g_N[h_N(x, j)] \psi_N[h_N(x, k), 0]$$

for  $k=1, 2, \dots, N$ , where  $h_N(x, p) = x$  and  $h_N(x, j) = h_N[h_N(x, j-1), 1]$  for  $j \geq 1$ .

Now letting  $k = [Nt]$  in (12) taking logarithms we obtain

$$\lim_{N \rightarrow \infty} \ln \psi_N(x, [Nt]) = -(1/2) x' \int_0^t x + ie^{-t} \chi_0 \cdot x \, dt,$$

which is the characteristic function of  $y(t)$ . The convergence of the f.d.d. is shown by a similar argument. To complete the proof of weak convergence a combination of the methods of Stone (1961) and Billingsley (1956) are used.

The method outlined above to show convergence of the f.d.d. can be used for a variety of related urn models, some of which are associated with queueing problems. These results will appear in future publications.

## 7. Weak Convergence of a Sequence of Quickest Detection Problems

Consider a production process which is in one of two states, a good state and a bad state, which correspond to being in control and out of control. Production begins with the process in control and after each item is produced there is a probability  $w$  of the process going out of control. A statistical control procedure is desired which will enable one to detect the fact that the process is out of control in some optimal manner. This model of a production process was first introduced by Girshick and Rubin (1952) and later discussions of the problem are due to Shiryaev (1963), Taylor (1967), and Bather (1967). Most of the work carried out in these papers deals with a continuous time analog in which a Brownian motion process with mean 0 has a drift of 1 introduced after some independent exponential time. The corresponding optimal statistical control procedures are then derived and proposed as good rules for controlling the discrete processes. The passage from discrete to continuous and back to discrete

has never been carried out in a rigorous way. It turns out that the notions of weak convergence are exactly what is required.

Our discussion, based on the paper Iglehart and Taylor (1967), begins with a sequence of truncated processes which can be easily described as follows. In the truncated problem of length  $n$  ( $\geq 2$ ) the process produces  $n$  independent items and goes out of control at a random time  $T_n$  ( $\leq n$ ). All items produced at or before time  $T_n$  are assumed to have a random quality with distribution function  $F_0$  (having mean 0 and variance 1). All items produced after  $T_n$  possess quality given by  $F_1(x) = F_0(x-1)$ ; i.e., the process when out of control shifts the d.f.  $F_0$  by 1 unit. The distribution of  $T_n$  is given by

$$\Pr(T_n = j) = \begin{cases} \frac{\tau}{n} (1 - \frac{\tau}{n})^{j-1} & , \quad j = 1, 2, \dots, n-1 \\ (1 - \frac{\tau}{n})^{n-1} & , \quad j = n \end{cases}$$

which is simply a geometric distribution with parameter  $\tau/n$  truncated at  $n-1$  with the remaining mass lumped at  $j=n$ . If  $T_n = n$ , then all  $n$  items produced have quality given by  $F_0$ . The process is just turned off after  $n$  items have been produced and if  $T_n = n$  it would be out of control, but this is then irrelevant.

We introduce two sequences of measures on  $\mathcal{C}$  (Borel sets of  $C[0,1]$ ) as follows. Let  $X_1, X_2, \dots$  be a sequence of independent random variables with d.f.  $F_0$  and define

$$X_n(k) = \left( \sum_{i=1}^k X_i \right) / n^{1/2}$$

for  $k = 1, 2, \dots, n$  and  $X_n(0) = 0$ . Then the paths  $x_n(t)$  in  $C[0,1]$  are obtained by setting  $x_n(t) = X_n(k)$  for  $t = k/n$ ,  $k = 0, 1, \dots, n$  and by linear interpolation for other values of  $t$ . Let  $\mu_n$  denote the measure induced on

by  $x_n(\cdot)$ . Next define the continuous paths  $(\theta_n(\cdot))$  on  $[0,1]$  by

$$\theta_n(t) = \begin{cases} 0 & t \leq T_n/n \\ (t - T_n/n) & t > T_n/n \end{cases}.$$

The measures induced by  $\theta_n(\cdot)$  we denote by  $\lambda_n$ . Clearly, the observed process of production corresponds to  $y_n(t) = x_n(t) + \theta_n(t)$ .

From the work of Prokhorov (1956) we know that  $\mu_n \Rightarrow \mu$ , where  $\mu$  is Wiener measure for paths starting at 0. We now introduce the measure  $\lambda$  on induced by the process  $\theta(\cdot)$  on  $[0,1]$  defined by

$$\theta(t) = \begin{cases} 0 & t \leq T \\ (t - T) & t > T \end{cases},$$

where the random variable  $T$  has an exponential density  $f_T(t)$  with parameter  $\tau$  for  $0 \leq t < 1$  and assumes the value 1 with probability  $e^{-\tau}$ . Using characteristic functions it is easy to show that the f.d.d. of  $\lambda_n$  converge to those of  $\lambda$ . Furthermore, since the support of all the measures  $(\lambda_n)$  is contained in the compact set  $M \subset C[0,1]$  given by

$$M = \{x: x(t)=0, t \leq t_0; x(t)=t-t_0, t > t_0; \text{ for some } t_0 \in [0,1]\},$$

it follows from Prokhorov (1956) that the family of measures is tight and thus  $\lambda_n \Rightarrow \lambda$ . Finally, since the measures induced by  $y_n(t)$  are simply the convolution  $\lambda_n * \mu_n$ , and additional argument shows that  $\lambda_n * \mu_n \Rightarrow \lambda * \mu$ .

Girshick and Rubin (1952) show that the optimal control (under a cost structure which we won't mention) is to stop the process when the posteriori probability that the process will be out-of-control for the next item produced,

given the history of observations, exceeds a certain level. It is more convenient to consider a monotone function of this posteriori probability which maps each path of the process  $\{y_n(t): 0 \leq t \leq 1\}$  into  $C[0,1]$ . This sequence of mappings (one for each  $n$ ) is continuous and converges uniformly on compact sets of  $C[0,1]$ . Another result of Prokhorov (1956) implies that the measures induced by these monotone functions converge weakly. Finally, this last result implies that the distribution of the optimal stopping times and the optimal costs converge. Thus we have established in a rigorous manner the relationship between the discrete models and the continuous analog.

#### 8. Other Work on Convergence of Processes in Applied Probability

In this final section we shall mention briefly some other work on convergence of processes. The area of applied probability in which diffusion approximations have been most widely used is population genetics. This work was initiated by Fisher and Wright in the 1930's. We have not discussed any of these applications since a comprehensive review is available by Kimura (1964). A subsequent paper which treats diffusion approximations in genetics from the point of view discussed here is Karlin and McGregor (1964).

In branching processes several papers have recently appeared which deal with convergence of processes. These papers are Lamperti (1967), Lamperti and Ney (1967), and Lamperti (1967).

Distribution-free statistics such as those of the Kolmogorov-Smirnov and Cramér-von-Mises types can be defined as functionals on the sequence of empirical stochastic processes. Convergence of these processes has been studied by Joob, Donsker, and Prokhorov. For an excellent summary of this work, complete references, and further extensions the reader should consult Pyke (1965). Two additional papers which deal with order statistics and related random variables are Dwass (1964) and Lamperti (1964).

## References

- Bather, J. (1967). On a quickest detection problem. Ann. Math. Stat. 38, 711-724.
- Billingsley, P. (1956). The invariance principle for dependent random variables. Trans. Amer. Math. Soc. 83, 250-268.
- Borovkov, A. (1964). Some limit theorems in the theory of mass service. Theor. Probability Appl. 9, 550-565. (English translation)
- Darling, D. and Siegert, A. (1953). The first passage problem for a continuous Markov process. Ann. Math. Stat. 24, 624-639.
- Donsker, M. (1951). An invariance principle for certain probability limit theorems. Mem. Amer. Math. Soc. No. 6.
- Dwass, M. (1964). Extremal processes. Ann. Math. Stat. 35, 1718-1725.
- Erdős, P. and Kac, M. (1946). On certain limit theorems in the theory of probability. Bull. Amer. Math. Soc. 52, 292-302.
- Feller, W. (1957). Generalized second order differential operators and their lateral conditions. Illinois J. Math. 1, 459-504.
- Girshick, M. and Rubin, H. (1952). A Bayes approach to a quality control model. Ann. Math. Stat. 23, 114-125.
- Iglehart, D. (1965). Limit diffusion approximations for the many server queue and the repairman problem. J. Appl. Prob. 2, 429-441.
- Iglehart, D. (1967). Limit theorems for the multi-urn Ehrenfest model. Technical Report No. 19, Department of Operations Research, Cornell University.
- Iglehart, D. and Taylor, H. (1967). Weak convergence for a sequence of quickest detection problems. Technical Report No. 30, Department of Operations Research, Cornell University.
- Itô, K. and McKean, H., Jr. (1965). Diffusion processes and their sample paths. Springer-Verlag, Berlin.
- Kac, M. (1947). Random walk and the theory of Brownian motion. Amer. Math. Monthly 54, 369-391.
- Karlin, S. and McGregor, J. (1958). Many server queueing processes with Poisson input and exponential service times. Pacific J. Math. 8, 87-118.
- Karlin, S. and McGregor, J. (1964). On some stochastic models in genetics. J. Gurland (ed.), Stochastic Models in Medicine and Biology. U. of Wisconsin Press, Madison, 245-279.
- Karlin, S. and McGregor, J. (1965). Ehrenfest urn models. J. Appl. Prob. 2, 352-376.
- Kimura, M. (1964). Diffusion models in population genetics. J. Appl. Prob. 1, 177-232.

Kingman, J. (1962). On queues in heavy traffic. J. Roy. Statist. Soc. Ser. B 24, 383-392.

Kingman, J. (1965). The heavy traffic approximation in the theory of queues. W. Smith and W. Wilkinson (eds.) Proc. of the Symp. on Congestion Theory, 137-159.

Knight, F. (1962). On the random walk and Brownian motion. Trans. Amer. Math. Soc. 103, 218-228.

Lalchandani, A. (1967). Some limit theorems in queueing theory. Technical Report No. 29, Department of Operations Research, Cornell University.

Lamperti, J. (1964). On extreme order statistics. Ann. Math. Stat. 35, 1726-1737.

Lamperti, J. (1967a). Limiting distributions for branching processes. Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability. (To appear.)

Lamperti, J. (1967b). The limit of a sequence of branching processes. Z. Wahrscheinlichkeitstheorie verw. Geb. 7, 271-288.

Lamperti, J. and Ney, P. (1967). Conditioned branching processes and their limiting diffusions. (To be published.)

Luchak, G. (1958). The continuous time solution of the equations of the single channel queue with a general class of service time distributions by the method of generating functions. J. Roy. Statist. Soc. Ser. B 20, 176-181.

Prabhu, N. (1965). Queues and Inventories. John Wiley & Sons, New York.

Prokhorov, Yu. (1956). Convergence of random processes and limit theorems in probability theory. Theor. Probability Appl. 1, 157-214. (English translation.)

Prokhorov, Yu. (1963). Transient phenomena in processes of mass service. Litovsk. Mat. Sb. 3, 199-206. (In Russian.)

Pyke, R. (1965). Spacings. J. Roy. Statist. Soc. B 27, 395-449.

Shiryaev, A. (1963). On optimum methods in quickest detection problems. Theor. Probability Appl. 8, 22-46. (English translation.)

Skorokhod, A. (1956). Limit theorems for stochastic processes. Theor. Probability Appl. 1, 262-290. (English translation.)

Stone, C. (1961). Limit theorems for birth and death processes and diffusion processes. Ph.D. thesis. Stanford University.

Stone, C. (1963). Limit theorems for random walks, birth and death processes, and diffusion processes. Illinois J. Math. 7, 638-660.

Taylor, H. (1967). Statistical control of a Gaussian process. Technometrics 9, 29-41.

Viskov, O. (1964). Two asymptotic formulas in the theory of queues. Theor. Probability Appl. 9, 158-159. (English translation.)

Viskov, O. and Prokhorov, Yu. (1964). The probability of loss calls in heavy traffic. Theor. Probability Appl. 9, 92-96. (English translation.)

**OPTIMAL STOCHASTIC CONTROL**

**by**

**HERMAN CHERNOFF**

**at the**

**American Mathematical Society Summer Seminar**

**on the**

**Mathematics of the Decision Sciences**

**Stanford University**

**July - August 1967**

# OPTIMAL STOCHASTIC CONTROL<sup>1/</sup>

Herman Chernoff

Stanford University

## 1. Introduction.

Certain techniques which were developed for sequential analysis have been found to apply to a wide variety of problems including some stochastic control problems. We shall outline these techniques and indicate how they may be applied. The main tool is the representation of incoming information in terms of continuous time Wiener process which relates these problems to the solution of free Boundary Problems involving the heat equation.

## 2. Some Problems.

Consider the following distinct problems.

Problem 1. (A Sequential Analysis Problem). The random variable  $x$  is normally<sup>2/</sup> distributed with unknown mean  $\mu$  and known variance  $\sigma^2$ . The

---

<sup>1/</sup> Prepared with the partial support of NSF Grant GP 5705. An amplified version of this paper will be submitted to Sankhyā.

<sup>2/</sup> We shall use the following notation throughout the paper. Let  $\varphi(x) = (2\pi)^{-1/2} e^{-x^2/2}$  and  $\Phi(x) = \int_{-\infty}^x \varphi(y) dy$  to represent the standard normal density and cumulative distribution functions. The normal distribution with mean  $\mu$  and variance  $\sigma^2$  is  $\mathcal{N}(\mu, \sigma^2)$  and has density  $n(x; \mu, \sigma^2) = \sigma^{-1} \varphi((x-\mu)/\sigma)$  and cdf  $\Phi[(x-\mu)/\sigma]$ . We denote the probability distribution (law) of a random variable  $X$  by  $\mathcal{L}(X)$  and its mathematical expectation by  $E(X)$ .  $\mathcal{L}(X|Y)$  and  $E(X|Y)$  represent the conditional distribution and expectation of  $X$  given  $Y$ .

statistician must decide whether  $\mu > 0$  or  $\mu < 0$  and the cost of an incorrect decision is  $k|\mu|$ ,  $k > 0$ . He is permitted to sample sequentially (one observation at a time) at a cost of  $c$  per observation. He may stop sampling at any time and make a decision. The total cost will be  $cn$  if the decision is correct and  $cn + k|\mu|$  if it is wrong where  $n$  is the number of observations taken. The cost is a random variable whose distribution depends on the unknown  $\mu$  and the (sequential) procedure used.

To determine an optimal procedure one must specify some criterion of optimality. It is convenient to treat this problem in a Bayesian context assuming that the unknown  $\mu$  is normally distributed with mean  $\mu_0$  and variance  $\sigma_0^2$  (both known). Then a sequential procedure determines an expected cost, and one may seek that procedure which minimizes the expected cost.

Problem 2. (A Stopping Problem). Let  $\{X_n, -m \leq n \leq 0\}$  be a stochastic process such that  $X_{-m} = x$  is specified and  $X_{n+1} = X_n + u_n$  where the  $u_i$  are independently and normally distributed with mean 0 and variance 1. Thus it is convenient to think of  $X_n$  changing as the subscript  $n$  increases to zero. For each  $n < 0$ , an observer can stop the process and collect  $X_n$  or he can wait till  $n = 0$  at which point he collects 0 if  $X_0 \geq 0$  and  $X_0^2$  if  $X_0 < 0$ . However, he must pay 1 for each observation. What constitutes an optimal procedure for stopping?

Problem 3. (Pocket Control-Infinite Fuel). A rocket is directed toward Mars. At certain time points  $\{t_i\}$ , instruments make measurements estimating the distance by which it will miss. The miss distance can

then be adjusted by an "instantaneous" use of fuel. The cost of missing by an amount  $y$  is  $ky^2$ . The cost per unit of fuel is  $c$ . As much fuel as is desired is available. How should fuel be allocated to minimize expected total cost?

Certain simplifications and specifications are necessary to make this problem meaningful and tractable. The assumption of an infinite supply of available fuel is already such a simplifying assumption. Let us replace the natural two-dimensional miss vector by a one dimensional value which can take on positive and negative values. Let us assume that if an amount of fuel  $\Delta$  is used at time  $t$  the miss distance is changed by  $\pm e(t)\Delta$  where  $e(t)$  represents the efficiency of fuel at time  $t$ . Since fuel is used to change direction, the efficiency  $e(t)$  is greatest at the beginning of the flight when the rocket is far away from the target. We shall assume that the amount of fuel used and its effect can be controlled and measured with precision. Let us now assume that the measurements estimating the miss distance are  $x_i$  at time  $t_i$  where the  $x_i$  are independent and normally distributed with mean equal to the miss distance (provided no additional fuel is used) and known variance equal to  $\sigma_i^2$ .

Here as in the Sequential Analysis problem, the performance characteristics of a procedure depend on the unknown value of a fundamental parameter. In the Sequential Analysis problem that was  $\mu$ . Here it is the unknown miss distance that would be obtained if no adjustments were made. As in the sequential analysis problem we find it convenient to assume that the unknown miss distance has a normal prior distribution with specified mean  $\mu_0$  and variance  $\sigma_0^2$ . Then there is an expected

cost for each procedure and the problem of selecting an optimal procedure is meaningful.

These three problems have more flavor if the competing factors are indicated qualitatively. In the sequential problem, after much data has been accumulated one is either reasonably certain of the sign of  $\mu$  or that  $|\mu|$  is so small that the loss of deciding wrong is less than the cost of another observation. Here one expects the proper procedure to be such that one stops and makes a decision when the current estimate of  $|\mu|$  is sufficiently large and continues sampling otherwise. What constitutes sufficiently large depends on, and should decrease with, the number of observations or equivalently the precision of the estimate. It can be shown that after a certain sample size it pays to stop no matter what the current estimate of  $|\mu|$  is.

In the Stopping Problem, it is clear that if  $X_n$  is sufficiently negative (depending on  $n$ ) one ought to pay the cost of continuing one more step. It is to be expected that there are limits  $y_n$  so that for  $X_n < y_n$  it pays to continue and for  $X_n \geq y_n$  it pays to stop.

In the rocket problem, when the rocket is close to target, fuel efficiency may be so low that even though the miss distance is practically known one would not use fuel unless the miss distance were very great. When the rocket is far from the target fuel is effective but the miss distance is not well known and so one is reluctant to make an adjustment for fear of overshooting or adjusting in the wrong direction. Here one should expect the solution to have the following property. There are limits  $y_1$  at time  $t_1$  such that if the estimated miss distance exceeds  $y_1$  in absolute value, enough fuel is used to make the adjusted estimate

to  $y_1$ . It seems reasonable to expect the  $y_i$  to be large at the beginning of flight and at the end of the flight and relatively small in between.

For specific values of the constants, all three of these problems can be solved numerically by the backward induction techniques of dynamic programming. For the sequential analysis problem care must be taken to initiate the backward induction at a sample size  $n$  sufficiently large so that no matter what the estimate of  $\mu$  is, the optimal procedure will lead to a decision rather than to additional sampling. The technique of the backward induction can be summarized by the equation

$$(2.1) \quad \rho_n(\xi_n) = \inf_{a_n} E \rho_{n+1}(\xi_{n+1}(a_n, \xi_n))$$

where  $\rho_n(\xi_n)$  is the expected cost of an optimal procedure given the history  $\xi_n$  up to stage  $n$ ,  $\xi_{n+1}(a_n, \xi_n)$  describes the history up to stage  $n+1$  which may be random with distribution depending on  $\xi_n$  and the action  $a_n$  taken at stage  $n$ . It is possible to show that in the sequential analysis and rocket control problems  $\xi_n$  is adequately summarized by the mean and variance of the posterior distribution of  $\mu$  while in the stopping problem  $X_n$  may be used to describe  $\xi_n$ . The minimizing  $a_n$  which depend on  $\xi_n$  determine the optimal procedure.

To illustrate, we note that for the stopping problem  $\rho_0(x) = m$  for  $x \geq 0$  and  $\rho_0(x) = x - x^2$  for  $x \leq 0$ . Since reaching  $n = 0$  implies a payment of  $m$  for continuing to stage  $n = -1$ , the choice of stopping at  $X_{-1} = x$  leads to a cost of  $x - x^2$  while the choice of continuing leads to  $X_0 = x + u$  and an expected cost of  $\int_{-\infty}^{\infty} \rho_0(x+u) \phi(u) du$ .

Thus

$$\rho_{-1}(x) = \min[(m-1), m - \int_{-\infty}^{-x} (x+u)^2 \varphi(u) du]$$

and the best action for  $X_{-1} = x$  is to stop or continue depending on which of the two terms in the brackets is smaller. Having evaluated  $\rho_{-1}(x)$ , one can in principle proceed in the same way to obtain  $\rho_{-2}$  and optimal decision for  $n = -2$  as a function of  $X_{-2}$ , etc.

The rocket and sequential analysis problem seem more complex in that they involve the posterior distributions, but the calculus of posterior distributions (to be discussed in Section 6) when dealing with normal random variables and normal priors permits these problems to be treated with equal facility.

In a sense then, these problems are trivial. If, however, it is desired to derive some overall view of how the solutions depend on the various parameters, the simple though extensive numerical calculations of the backward induction are not adequate.

An approach which seems to have particular relevance to large sample theory is that of replacing the discrete time random variables by analogous continuous time stochastic processes. The use of the Wiener process seems especially relevant and serves to convert the problem to one in which the analytic methods of partial differential equations can serve fruitfully.

### 5. Wiener Process.

Suppose  $x_1, x_2, \dots, x_n$  are independently and identically distributed with mean  $\mu$  and variance  $\sigma^2$ . Let  $X_n = x_1 + x_2 + \dots + x_n$ . Then for

$n \geq m \geq 0$ ,  $X_n - X_m$  is independent of  $X_1, X_2, \dots, X_m$ , and normally distributed with mean  $(n-m)\mu$  and variance  $(n-m)\sigma^2$ . When  $n$  is large, a graph of the discrete  $X_n$  process resembles the analogous, continuous time process  $\{X(t): 0 \leq t\}$  which has the following properties. The function  $X(t)$  is continuous,  $X(0) = 0$ , and for  $0 \leq t_1 \leq t_2$ ,  $X(t_2) - X(t_1)$  is independent of  $\{X(t): 0 \leq t \leq t_1\}$  and is normally distributed with mean  $\mu(t_2 - t_1)$  and variance  $\sigma^2(t_2 - t_1)$ . This may be referred to as a Gaussian process with independent increments or as a Wiener process with drift  $\mu$  and variance  $\sigma^2$  per unit time. Typically the drift is not referred to when we consider the case  $\mu = 0$ . There is a trivial but occasionally useful variation where  $X(t)$  is initiated at the point  $X(t_0) = x_0$  rather than at  $X(0) = 0$ . Notationally it is convenient to refer to the process by the equations

$$\begin{aligned}
 E[dX(t)] &= \mu dt \\
 \text{Var}[dX(t)] &= \sigma^2 dt
 \end{aligned}
 \tag{3.1}$$

This is especially convenient for variations of the Wiener process which are derived by changing the scales. Observe that if  $E[dX(t)] = 0$  and  $\text{Var}[dX(t)] = dt$ , the transformation  $t^* = a^2 t$ ,  $X^*(t^*) = aX(t)$  yields  $E[dX^*(t^*)] = 0$  and  $\text{Var}[dX^*(t^*)] = dt^*$ .

Although the stopping problem (Problem 2, Section 2) does not involve an unknown parameter  $\mu$ , an analogue of this problem can be posed in terms of a continuous time Wiener process without drift originating from a given point  $(x, t)$ ,  $t \leq 0$ .

#### 4. Posterior Distributions.

Inasmuch as the Sequential Analysis and Rocket Control problems involve unknown parameters which may be estimated by incoming data, both have a statistical component. For statistical problems in a Bayesian context, the posterior distribution of the unknown parameter is crucial.

In the discrete case suppose that  $\mu$  has prior distribution  $\eta(\mu_0, \sigma_0^2)$  (normal with mean  $\mu_0$  and variance  $\sigma_0^2$ ), and for given  $\mu$ , the data  $x_1, x_2, \dots, x_n$  are independent with distribution laws  $\mathcal{L}(x_i) = \eta(\mu, \sigma_i^2)$ , and  $\sigma_i^2$  known. Then the posterior distribution of  $\mu$  given the data is (see [18]).

$$(4.1) \quad \mathcal{L}(\mu | x_1, \dots, x_n) = \eta(Y_n, s_n)$$

where

$$(4.2) \quad Y_n = (\mu_0 \sigma_0^{-2} + x_1 \sigma_1^{-2} + \dots + x_n \sigma_n^{-2}) / (\sigma_0^{-2} + \sigma_1^{-2} + \dots + \sigma_n^{-2}), \quad n \geq 0$$

and

$$(4.3) \quad s_n^{-1} = \sigma_0^{-2} + \sigma_1^{-2} + \dots + \sigma_n^{-2}, \quad n \geq 0$$

Here  $Y_n$ , the mean of the posterior distribution, may be called the (posterior) Bayes Estimate of  $\mu$ . It is a weighted average of the individual estimates  $x_i$  weighted by the precisions  $(\sigma_i^{-2})$  where the prior distribution is treated as an estimate with mean  $\mu_0$  and precision  $\sigma_0^{-2}$ . Similarly  $Y_n$  may be regarded as a summary of the previous information and as an estimate of  $\mu$  with precision  $s_n^{-1}$ .

Since  $Y_n$  is the Bayes estimate of  $\mu$ , one should expect that for

$n > m$ ,  $E(Y_n | Y_m) = Y_m$ . It is somewhat more surprising to find by routine but tedious calculations that

$$(4.4) \quad \mathcal{L}(Y_n - Y_m | Y_m) = \eta(0, s_m - s_n) \quad n \geq m \geq 0.$$

## 5. Continuous Time Problems.

The results of section 4 are of consequence in the continuous time analogue of the sequential analysis problem which we now state.

Problem 1\*. (Sequential Analysis, Continuous Time). Find an optimal procedure for testing  $H_1: \mu \geq 0$  vs.  $H_2: \mu < 0$  when the cost of an incorrect decision is  $k|\mu|$ , the cost of sampling is  $c$  per unit time, and the data consists of a Wiener process  $X(t)$  unknown drift  $\mu$  and known variance  $\sigma^2$  per unit time. The unknown value of  $\mu$  has prior distribution  $\mathcal{L}(\mu) = \eta(\mu_0, \sigma_0^2)$ .

As a consequence of the results of section 4 we have for Problem 1\*, the posterior distribution of  $\mu$  given by

$$(5.1) \quad \mathcal{L}(\mu | X(t'), 0 \leq t' \leq t) = \eta(Y(s), s)$$

where

$$(5.2) \quad Y(s) = [\mu_0 \sigma_0^{-2} + X(t) \sigma^{-2}] / (\sigma_0^{-2} + t \sigma^{-2}),$$

$$(5.3) \quad s = \sigma_0^{-2} + t \sigma^{-2},$$

and  $Y(s)$  is a Wiener process in the  $-s$  scale, originating at  $(y_0, s_0) = (\mu_0, \sigma_0^2)$ , i.e.

$$(5.4) \quad E[dY(s)] = 0, \quad \text{Var}\{dY(s)\} = -ds.$$

Note that  $s$  decreases from  $s_0 = \sigma_0^2$  as information accumulates.

Since the  $X$  process can be recovered from the  $Y$  process it suffices to deal with the latter which measures the current estimate of  $\mu$  and which is easier to analyze.

In the sequential problem, when the statistician stops sampling he must decide between  $H_1: \mu \geq 0$  and  $H_2: \mu < 0$ . The posterior expected cost associated with deciding in favor of  $H_1$  at time  $t$  (when  $Y(s) = y$ ) is then

$$\int_{-\infty}^0 k|\mu|n(\mu; y, s) d\mu$$

This quantity is readily computed and found to be  $k\sqrt{s}\psi^+(y/\sqrt{s})$  where  $\psi^+(u) = \phi(u) - u[1 - \Phi(u)]$ . Similarly the posterior expected cost associated with deciding  $\mu < 0$  is  $k\sqrt{s}\psi^-(u)$  where  $\psi^-(u) = \phi(u) + u\Phi(u)$ . It is easy to see that if sampling is stopped at  $Y(s) = y$  the decision should be made on the basis of the sign of  $y$  and the expected cost of deciding plus the cost of sampling is given by

$$(5.5) \quad d(y, s) = c\sigma^2 s^{-1} + k\sqrt{s}\psi(y/\sqrt{s}) - c\sigma^2/\sigma_0^2$$

where

$$(5.6) \quad \psi(u) = \begin{cases} \phi(u) - u[1 - \Phi(u)] & u \geq 0 \\ \phi(u) + u\Phi(u) & u < 0 \end{cases}$$

Thus the continuous time sequential analysis problem may be regarded simply as the following stopping problem. The Wiener process  $Y(s)$  is observed. The statistician may stop at any value of  $s > 0$  and pay

$d(Y(s), s)$ . Find the stopping procedure which minimizes the expected cost. In this version of the problem using the posterior Bayes Estimate, the statistical aspects involving the unknown parameter  $\mu$  have been abstracted.

The original discrete time sequential problem can also be described in terms of this stopping problem by adding the proviso that allowable stopping values of  $s$  are restricted to  $s_0, s_1, s_2, \dots$ , where  $s_n = (\sigma_0^{-2} + n\sigma^{-2})^{-1}$ . At this point it should be reasonably straightforward for the reader to see that the discrete version can be treated numerically by backward induction in terms of the  $Y(s)$  process starting from  $s_n \leq c^2/k^2\psi^2(0) = 2\pi c^2/k^2$ .

We now present the continuous time version of the stopping Problem 2 of Section 2.

Problem 2\*. (Stopping Problem). Let  $Y(\Delta)$  be a Wiener process in the  $-s$  scale with  $E\{dY(s)\} = 0$ ,  $\text{Var}\{dY(s)\} = -ds$ , originating from  $(y_0, s_0)$ . Let

$$(5.7) \quad d(y, s) = \begin{cases} -y^2 & \text{if } y < 0 \text{ and } s = 0 \\ -s & \text{otherwise } (s \geq 0) \end{cases}$$

Find the stopping time to minimize  $E\{d(Y(s), s)\}$ .

A more literal translation of Problem 2 of Section 2 would yield a stopping cost  $d^*(y, s) = d(y, s) + s_0$ . Since the difference is constant it does not affect the solution.

While the rules of computing posterior distributions extend to the rocket control problem, that problem is not trivially reduced to a

stopping problem. However, we shall see that the continuous version of the infinite fuel problem posed in Section 2 is also equivalent to a related stopping problem.

#### 6. Continuous Time Stopping Problems: Relevance of Stopping Sets.

A general class of stopping problems may be described as follows. Let  $Y(s)$  be a Wiener process in the  $-s$  scale originating from  $(y_0, s_0)$  with  $E(dY(s)) = 0$  and  $\text{Var}(dY(s)) = -ds$ . Let  $d(y, s)$  be a specified stopping cost. Select a stopping procedure  $S$  to minimize the risk

$$(6.1) \quad b(y_0, s_0) = E(d(Y(S), S)) .$$

A stopping procedure  $S$  is a measurable rule which determines the stopping time  $S$  in terms of the "past history" of  $Y(s)$ . Technically, in measure theoretic terms this may be translated to mean

$$(6.2) \quad (S \geq s_1) \in \mathcal{B}(Y(s); s_0 \geq s \geq s_1)$$

where the right hand side is the Borel Field generated by the process from  $s_0$  to  $s_1$ . Stopping procedures may be subjected to restrictions which are either of the form that stopping is not allowed on certain sets of points  $(y, s)$  or that stopping is automatic on other sets. For example in the continuous time versions of both problems 1 and 2 of Section 2, stopping must take place if  $s = 0$ . In a trivial sense the discrete time version of the sequential analysis problem may be regarded as a continuous time problem where stopping is not permitted except at

a certain set of values of  $s$ .

While the discrete time problems of Section 2 are theoretically trivial insofar as the solutions can be computed by backward induction this is not the case for the continuous time problems. Even discrete time problems with an infinite sequence of possible decision times lead to difficulties. The problems of the existence and characterization of solutions are deep and much remains to be done to obtain precise rigorous results for the continuous time problem. We shall proceed in a heuristic fashion conveniently ignoring some of the more delicate questions which have to be faced ultimately.

Let

$$(6.3) \quad \rho(y_0, s_0) = \inf b(y_0, s_0)$$

among all procedures  $S$ . Note that  $\rho(y_0, s_0) \leq d(y_0, s_0)$ . Since  $Y$  is a process of independent increments, it follows that  $\rho(y, s)$  also represents the best that can be expected once  $Y(s) = y$  is reached, irrespective of how it was reached. Then, a characterization of an optimal procedure (under regularity conditions) is described by

$$(I) \quad S_0: \text{"Stop as soon as } \rho(Y(s), s) = d(Y(s), s)\text{"}$$

Since the optimal procedure  $S_0$  is characterized by the continuation set

$$(6.4) \quad G_0 = \{(y, s): \rho(y, s) < d(y, s)\}$$

and the stopping set

$$(6.5) \quad \mathcal{L}_0 = \{(y, s): \rho(y, s) = d(y, s)\}$$

we shall restrict our attention to procedures which can be represented by a continuation set  $\mathcal{C}$  or its complement the stopping set  $\mathcal{S}$ .

It is interesting to note that the characterization (I) does not depend on the initial point  $(y_0, s_0)$  and thus it yields the solution for all initial points simultaneously, minimizing  $b(y, s)$  uniformly for all  $(y, s)$ .

Under suitable regularity conditions on  $d(y, s)$ , the solution of the continuous time stopping problems may be approximated by discrete time versions corresponding to a finite sequence of permitted stopping times  $(s_1, s_2, \dots, s_n)$ . Since a discrete version permits less choice, the corresponding optimal risk  $\rho^*$  is larger and the corresponding optimal continuation set  $\mathcal{C}_0^*$  intersects  $s = s_1$  on a smaller set than does  $\mathcal{C}_0$ . As more elements are adjoined to the set of permitted stopping times,  $\rho^*$  decreases and the set where  $\mathcal{C}_0^*$  intersects  $s = s_1$  increases. In this way  $\rho$  and  $\mathcal{C}_0$  may be derived as limits of monotone sequences.

#### 7. Stopping Problems. Relevance of Heat Equation.

The Wiener process is intimately related to the heat equation. Suppose, for example that  $b(y, s)$  is the expected cost corresponding to an open continuation set  $\mathcal{C}$  and stopping cost  $d(y, s)$ . Then we shall demonstrate that

$$(7.1) \quad \frac{1}{2} b_{yy}(y, s) = b_s(y, s) \quad (y, s) \in \mathcal{C}$$

while

$$(7.2) \quad b(y, s) = d(y, s) \quad (y, s) \in \mathcal{S}.$$

Suppose  $(y, s) \in \mathcal{C}$ . Then, the probability of stopping between  $s+\delta$  and

$s$  is  $O(\delta)$  and  $Y$  changes from  $Y(s+\delta)$  to  $Y(s)$ . Consequently

$$\begin{aligned} b(y, s+\delta) &= E\{b(Y(s), s) | Y(s+\delta) = y\} + O(\delta) \\ (7.3) \quad &= E\{b(y+w\sqrt{\delta}, s)\} + O(\delta) \end{aligned}$$

where we use  $w$  as a generic  $\mathcal{N}(0,1)$  random variable

$$\begin{aligned} b(y, s+\delta) &= E\{b(y, s) + w\sqrt{\delta} b_y(y, s) + \frac{1}{2} w^2(\delta) b_{yy}(y, s) + \dots\} + O(\delta) \\ &= b(y, s) + \frac{1}{2} b_{yy}(y, s)(\delta) + O(\delta) \end{aligned}$$

and

$$b_s = \frac{1}{2} b_{yy}$$

Doob has elaborated on the relationship between the Wiener process and the heat equation indicating that it represents the natural way in which to study the heat equation. To digress briefly and omitting regularity conditions, a subparabolic function  $u$  on an open set  $D$  is such that for a Wiener process  $Y(s)$  originating at  $(y, s)$

$$(7.4) \quad u(y, s) \leq E\{u(Y(S), S)\}$$

where  $S$  is the time when  $Y(s)$  first hits the boundary of an open set  $G \subset D$  of which  $(y, s)$  is an interior point. A parabolic function is one for which the inequality is replaced by equality. If the second derivatives are continuous then

$$(7.5) \quad \frac{1}{2} u_{yy} \geq u_s$$

for subparabolic functions with equality for parabolic functions. Thus solutions of the heat equation are identified with parabolic functions. To solve the Dirichlet Problem (solution of  $\frac{1}{2} u_{yy} = u_s$  in  $G$  subject to  $u = f$  on the boundary of  $G$ ) Doob takes  $u(y,s) = E\{f(Y(S),S)\}$ .

The concept of subparabolic functions provides another characterization of the optimal risk. We observe that

(II)  $\rho(y,s)$  is the maximal subparabolic function which is less than or equal to  $d(y,s)$ .

To see that  $\rho$  is subparabolic, take an arbitrary set  $G$  of which  $(y,s)$  is an interior point. Then  $E\{\rho(Y(S),S)\}$  represents the risk associated with the suboptimal procedure which does not stop as long as  $(Y(s),s) \in G$  but which proceeds optimally thereafter. Thus

$$(7.6) \quad \rho(y,s) \leq E\{\rho(Y(S),S)\}$$

and  $\rho$  is subparabolic. Let  $\rho_1$  be any subparabolic function such that  $\rho_1 \leq d$ . Using the optimal continuation set  $G_0$  for  $G$ , we have, for  $(y,s) \in G_0$ ,

$$\rho(y,s) = E\{d(Y(S_0),S_0)\} \geq E\{\rho_1(Y(S_0),S_0)\} \geq \rho_1(y,s) .$$

If  $(y,s) \notin G_0$ ,  $\rho(y,s) = d(y,s) \geq \rho_1(y,s)$  which completes the proof.

Given a function  $u(y,s)$  and a continuation set  $G$ , can we determine whether  $(u,G) = (\rho,G_0)$  i.e. whether  $(u,G)$  solve the optimization problem associated with the stopping problem. A sufficient condition is the following.

(III) If  $u \leq d$  is a subparabolic function which is parabolic on the open continuation set  $G$  and  $u = d$  elsewhere, then  $(u, G) = (\rho, G_0)$ , the solution of the optimization problem.

To show this, note that since  $u \leq d$  is subparabolic,  $u \leq \rho$ . But  $u$  is the risk corresponding to the continuation set  $G$ . Hence  $u \geq \rho$ .

#### 8. Stopping Problems - Free Boundary Problem.

Associated with a procedure described by a continuation set  $G$  we have a risk function  $b(y, s)$  which satisfies the heat equation in  $G$  subject to the boundary condition  $b = d$ . The solution of the optimal stopping problem minimizes  $b$  everywhere. Now we present the characterization

(IV)  $\rho_y(y, s) = d_y(y, s)$  on the boundary of  $G_0$ .

While the Dirichlet problem of finding  $b$  which satisfies the heat equation in a given  $G$  subject to  $b = d$  on the boundary is referred to as a boundary value problem, that of finding  $b$  and  $G$  so that  $b_y = d_y$  on the boundary also, is referred to as a Stefan or free boundary problem (f.b.p.). Property (IV) states that the solution of the optimization problem is the solution of the (f.b.p.).

To demonstrate (IV), let us assume that  $(y_0, s_0)$  is a point on a portion of the boundary above which are stopping points and below which are continuation points and that  $d_y$  exists at  $(y_0, s_0)$ . Then since  $\rho(y, s_0) = d(y, s_0)$  for  $y > y_0$  the right hand derivative  $\rho_y^+(y_0, s_0) = d_y(y_0, s_0)$ . For  $y < y_0$ ,  $\rho(y_0, s_0) \leq d(y_0, s_0)$  and hence

$\rho_y^-(y_0, s_0) \leq d_y(y_0, s_0)$ . Now we note that

$$(8.1) \quad \rho(y_0, s_0 + \delta) \leq E\{\rho(y_0 + w\sqrt{\delta}, s_0)\}$$

since the right hand side corresponds to the risk of the suboptimal procedure where one insists on sampling from  $s_0 + \delta$  to  $s_0$  and proceeding optimally thereafter. But

$$(8.2) \quad \begin{aligned} \rho(y_0 + w\sqrt{\delta}, s_0) &= \rho(y_0, s_0) + w\sqrt{\delta} \rho_y^+(y_0, s_0) + o(\sqrt{\delta}) \quad w > 0 \\ &= \rho(y_0, s_0) + w\sqrt{\delta} \rho_y^-(y_0, s_0) + o(\sqrt{\delta}) \quad w < 0 \end{aligned}$$

$$\begin{aligned} E\{\rho(y_0 + w\sqrt{\delta}, s_0)\} &= \rho(y_0, s_0) + \sqrt{\delta} (d_y \int_0^\infty w\phi(w)dw + \rho_y^- \int_{-\infty}^0 w\phi(w)dw) + o(\sqrt{\delta}) \\ &= \rho(y_0, s_0) + \sqrt{\delta} (d_y - \rho_y^-) + o(\sqrt{\delta}) \end{aligned}$$

Thus

$$\rho(y_0, s_0 + \delta) - \rho(y_0, s_0) \leq \sqrt{\delta} (d_y - \rho_y^-) + o(\sqrt{\delta})$$

Assuming that the difference quotient  $[\rho(y_0, s_0 + \delta) - \rho(y_0, s_0)](\delta)^{-1}$  is bounded it follows that  $d_y - \rho_y^- \geq 0$  which combined with the preceding results gives  $\rho_y = d_y$  on the boundary which establishes (IV).

Returning to the free boundary problem (f.b.p.) the following question arises. Is a solution of the f.b.p. necessarily a solution of the optimization problem? The answer is yes provided certain additional conditions are satisfied. That additional conditions are required is clear from the following considerations. Suppose that  $(u, \xi)$  is a

1

solution of both the f.b.p. ( $\frac{1}{2} u_{yy} = u_s$  on  $\mathcal{C}$ ,  $u = d$  and  $u_y = d_y$  on the boundary) and the optimization problem ( $u = \rho$ ,  $\mathcal{C} = \mathcal{C}_0$ ). If the problem is modified by sharply decreasing  $d$  below  $u$  on part of  $\mathcal{C}_0$ , then  $(u, \mathcal{C})$  remains a solution of the free boundary problem but the solution of the optimization problem changes. If  $d$  is sharply decreased on a small part of the stopping set near the boundary of  $\mathcal{C}$ , the optimal continuation region should be enlarged but here again  $(u, \mathcal{C})$  remains a solution of the free boundary problem.

These examples lead to sufficient conditions which are related to III. One of these (V) may be paraphrased to state that if one can't trivially improve on  $u$  (as was possible in the above counterexamples) then  $u = \rho$ . Let

$$(8.3) \quad h(y, s; s') = E[h(y + w\sqrt{s-s'}, s')] \quad s \geq s'$$

- (V) If  $u$  is the risk corresponding to the continuation set  $\mathcal{C}$  and  
(i)  $u(y, s) \leq d(y, s)$  and  
(ii)  $u(y, s; s') \geq d(y, s)$  for  $(y, s) \in \mathcal{C}$   
then  $(u, \mathcal{C})$  solves the optimization problem.

While (V) does not invoke the f.b.p. condition, that condition can be used to prove condition (ii) of (V). This yields

- (VI) If  $(u, \mathcal{C})$  is a solution of the f.b.p. where  $\mathcal{C}$  is a continuation set and  $u$  and  $d$  have bounded derivatives up to third order  
and  
(i)  $u(y, s) \leq d(y, s)$  and

$$(ii) \quad \frac{1}{2} d_{yy} > d_s \quad \text{on}$$

then  $(u, \mathcal{C})$  is a solution of the optimization problem.

In some applications (VI) is not enough because some of the conditions break down as  $s$  approaches its lower limit  $s_0$  (possibly  $-\infty$ ). In that case it suffices to invoke some supplementary condition which implies

$$(8.4) \quad \sup_{s \rightarrow s_0} |u(y, s) - \rho(y, s)| \rightarrow 0$$

#### 9. Solutions, Bounds and Expansions for Stopping Problems.

In the continuous version of Problem 2, we have a stopping problem which may be represented by

$$(9.1) \quad \begin{aligned} d(y, s) &= -s \quad \text{for } s > 0 \text{ and for } y \geq 0, s = 0 \\ d(y, s) &= -y^2 - s \quad \text{for } y \leq 0, s = 0. \end{aligned}$$

This problem has the trivial solution where  $\mathcal{C}_0 = \{(y, s): y < 0, s > 0\}$  and

$$(9.2) \quad \begin{aligned} \rho(y, s) &= -s \quad \text{for } y \geq 0, s \geq 0 \\ \rho(y, s) &= -y^2 - s \quad \text{for } y < 0, s > 0. \end{aligned}$$

The pair  $(\rho, \mathcal{C}_0)$  is a solution of the (f.b.p.) since  $\rho = d$  and  $\rho_y = d_y$  for  $y = 0$ . Property (V), Section 8 applies as does a modified version of Property (VI), Section 8 (a modification is required because  $\rho$  and  $d$  are not bounded).

Generally, stopping problems are not so easily solved. It is useful

to derive bounds on  $\rho$  and  $\mathcal{C}_0$ . To illustrate let us introduce a new stopping problem, the importance of which will be discussed later.

Problem 4. A stopping problem involving  $Y(s)$ ,  $EdY(s) = 0$ ,  $\text{Var } dY(s) = -ds$ ,

$$(9.3) \quad d(y,s) = \begin{cases} y & \text{for } s = 0, y \geq 0 \\ s^{-1} & \text{for } s > 0, y > 0 \\ 0 & \text{for } s \geq 0, y = 0 \end{cases}$$

and stopping is enforced when  $Y(s) = 0$  or  $s = 0$ .

Note that if  $s$  is large, the chances of obtaining  $Y(s) = 0$  (and zero cost) before  $s = 0$  is large and so one is encouraged to continue unless  $Y$  is large. If  $s$  is small, the cost of stopping,  $(s^{-1})$  is large compared to the cost of waiting till  $s = 0$  (approximately  $Y$ ) unless  $Y$  is large and one is encouraged to continue unless  $Y$  is large. Thus one expects  $\mathcal{C}_0$  to have a boundary which is high for  $s$  large and  $s$  small.

Let  $u(y,s)$  be an arbitrary solution of the heat equation. Let  $\mathcal{B}$  be the set on which  $u(y,s) = d(y,s)$ . If  $\mathcal{B}$  is the boundary of a continuation set  $\mathcal{C}$  the risk for the procedure defined by the continuation set  $\mathcal{C}$  is  $b(y,s) = u(y,s)$  on  $\mathcal{C}$  and  $b(y,s) = d(y,s)$  on  $\mathcal{B}$ . But then  $\rho(y,s) \leq b(y,s)$ . Thus if  $(y_0, s_0)$  is a point of  $\mathcal{C}$  where  $u < d$ , then  $\rho(y_0, s_0) < d(y_0, s_0)$  and  $(y_0, s_0)$  is a continuation point for the optimal procedure.

For Problem 4 take  $u_1(y,s) = y$  which is a solution of the heat equation.  $\mathcal{C} = \{(y,s): 0 < y < s^{-1}, s > 0\}$ . Since  $u_1(y,s) < s^{-1}$  at every point of  $\mathcal{C}$ ,  $\mathcal{C} \subset \mathcal{C}_0$  and the

boundary  $y_1(s) = s^{-1}$  is a lower bound for the optimal boundary  $\tilde{y}(s)$ ,  
i.e.

$$(9.4) \quad \tilde{y}(s) \geq y_1(s) = s^{-1} .$$

We now describe a method of finding upper bounds for the optimal boundary. In more general context these represent outer bounds for  $\mathcal{C}_0$ . Let  $u(y,s)$  be a solution of the heat equation. Let  $\mathcal{C}$  be the set on which  $u_y(y,s) = d_y(y,s)$ . Let  $\mathcal{C}$  be the continuation set for which  $\mathcal{B}$  is the boundary. If  $u \neq d$  on  $\mathcal{B}$  let  $h(s) = u(y,s) - d(y,s)$  along the boundary  $\mathcal{B}$  and let  $d^*(y,s) = d(y,s) + h(s)$ . Then  $(u, \mathcal{C})$  is a solution of the f.b.p. for  $d^*(y,s)$ . Suppose that  $(u, \mathcal{C})$  is also a solution of the optimality problem for  $d^*$  and  $h(s) \leq 0$  for  $s < s_2$  and  $h(s_2) = 0$ . Then the modified problem is a more "advantageous" problem than the original for  $s = s_2$  and

$$\rho(y, s_2) \geq u(y, s_2) .$$

If  $(y_2, s_2)$  is a stopping point for the modified problem

$$(9.5) \quad \rho(y_2, s_2) \geq u(y_2, s_2) = d^*(y_2, s_2) = d(y_2, s_2)$$

and  $(y_2, s_2)$  is a stopping point for the original problem.

In review we obtain outer bounds on the continuation set by finding arbitrary solutions of the heat equation which are suitable (i.e.  $u-d \leq 0$  along the boundary where  $u_y = d_y$ ). In principle this method is as elementary as the other method but in application it is usually more delicate.

To illustrate

$$u_2(y,s) = y - Be^{\frac{1}{2}a^2s} \sinh ay$$

is a solution of the heat equation for which  $u_{2y} = d_y = 0$  when  $y = y_2(s)$  which is determined by

$$1 - Bae^{\frac{1}{2}a^2s} \cosh ay = 0$$

For  $s = 0$ ,  $u_2 - d \leq 0$ . Along the boundary  $y = y_2(s)$ ,  $u_2 - d$  take on negative values for small positive  $s$ . The smallest positive value  $s_2$  of  $s$ , (if any), where  $u_2 - d$  vanishes is described by

$$y - Be^{\frac{1}{2}a^2s} \sinh ay = s^{-1}$$

Any pair of parameters  $(a,B)$  which yields such a pair  $(y_2, s_2)$  may be used and the corresponding point  $(y_2, s_2)$  is a point of  $\mathcal{J}_0$ . To find the best such point for a given  $s_2$ , we select  $a$  and  $B$  to minimize  $y_2$ . If for fixed  $(a,B)$ ,  $\frac{\partial(u_2-d)}{\partial s} > 0$  at  $(y_2, s_2)$  it would be possible to adjust  $a$  and  $B$  to decrease  $y_2$ . Thus we impose the third condition  $\frac{\partial(u_2-d)}{\partial s} = 0$

$$\frac{Ba^2}{2} e^{\frac{1}{2}a^2s} \cosh ay = s^{-2}$$

Together the three conditions lead to the representation for an outer bound  $\tilde{y}_2(s)$  for the boundary, i.e.  $\tilde{y}_2(s) \geq \tilde{y}(s)$  where  $y_2(s)$  satisfies

$$(9.6) \quad \frac{2^{1/2}}{s} (\tilde{y}_2 - s^{-1})^{1/2} = \tanh \left\{ \frac{2^{1/2} \tilde{y}_2}{s} (\tilde{y}_2 - s^{-1})^{-1/2} \right\}$$

It is of interest that  $s^{-1} < y_2(s) \leq s^{-1} + \frac{1}{2} s^2$  which indicates that  $\tilde{y}$  is well approximated by  $s^{-1}$  for small  $s$ . For large  $s$  better approximations are obtained by using similar arguments with solutions of the heat equation of the form

$$u = A\varphi\left(\frac{y}{\sqrt{s+h}}\right) \frac{y}{(\sqrt{s+h})^{3/2}}, \quad h \geq 0$$

which yield the lower bound

$$(9.7) \quad \tilde{y}_3(s) = \sqrt{s} \leq \tilde{y}(s)$$

and the upper bound

$$(9.8) \quad \tilde{y}_4(s) = (s+h)^{1/2} \geq \tilde{y}(s)$$

where  $h > e^{1/3}$  satisfies  $s = e^{1/2} h(h^{3/2} - e^{1/2})^{-1}$ . Together these show that  $\tilde{y}(s) = s^{1/2} \{1 + O(s^{-1})\}$  for large  $s$ .

Another but related approach to approximating the optimal boundary consists of finding asymptotic expansions for the risk and boundary near distinguished points of  $s$ ; these distinguished points are typically the end points of the range of interest. For example  $s = 0$  and  $\infty$  are important in examples 1, 2, and 4. The proofs that the formal expansions derived by methods to be briefly described do indeed represent approximations to the desired solution depend on arguments of the type described above.

One important class of solutions of the heat equation used in generating expansions is that generated by "sources of heat" along a vertical ( $s = \text{constant}$ ) line. Thus

$$(9.9) \quad u_0(y, s) = s^{-1/2} \varphi(\alpha) \quad , \quad \alpha = y/\sqrt{s}$$

represents a point source of heat at  $(y, s) = (0, 0)$  and yields a solution of the heat equation for  $s > 0$ . Similarly, functions of the form

$$u(y, s) = \int \frac{1}{\sqrt{s}} \varphi\left(\frac{y-y'}{\sqrt{s}}\right) h(y') dy' = \int h(y+w\sqrt{s}) \varphi(w) dw$$

satisfy the heat equation.

Such techniques lead to asymptotic expansions for the optimal solution of the sequential analysis problem (Problem 1\* of Section 5) of the form

$$\tilde{\alpha}(s) = \tilde{y}(s) s^{-1/2} \sim \{\log a^2 s^3 - \log 8\pi - 6(\log a^2 s^3)^{-1} + \dots\} \text{ as } s \rightarrow \infty$$

$$\tilde{\alpha}(s) = \tilde{y}(s) s^{-1/2} \sim \frac{1}{4} a s^{3/2} \left\{1 - \frac{a^2 s^3}{12} + \frac{7}{15 \cdot 16} a^4 s^6 - \dots\right\} \text{ as } s \rightarrow 0$$

where  $a = k/c$ .

#### 10. Control Problem.

Let us return to the rocket control problem (Problem 3, Section 2). For reasons to be discussed later an important case can be described in its continuous time version as follows.

Problem 3\*. One observes  $Y(s)$ , a Wiener process in the  $-s$  scale originating at  $(y_0, s_0)$  with  $E[dY(s)] = 0$  and  $\text{Var}[dY(s)] = -ds$ . As  $s$  decreases to 0,  $Y(s)$  may be adjusted instantaneously at any  $s > 0$  by an amount  $\Delta$  at a cost of  $|\Delta|d(s)$  where  $d(s) = s^{-1}$ . In addition to the accumulated cost due to the adjustments of  $Y(s)$ , there is a cost of

$\frac{1}{2} Y^2(0)$ . Find the rule for adjusting  $Y(s)$  which minimizes the expected cost.

The discrete time version of this problem corresponds to the specification of a finite set of  $s_1, s_0 \geq s_1 \geq s_2 \geq \dots \geq s_n = 0$  at which changes (corresponding to the use of fuel) are permitted. Let  $\rho^*(y, s_1)$  represent the expected additional cost associated with the optimal procedure for the discrete time version given  $Y(s_1) = y_1$ . Since it is possible to change  $y$  instantaneously at a cost of  $d(s_1)$  per unit  $y$

$$\rho^*(y, s_1) \leq \rho^*(y', s_1) + d(s_1)|y' - y| \quad .$$

from which it follows that

$$(10.1) \quad \left| \frac{\partial \rho^*(y, s_1)}{\partial y} \right| \leq d(s_1)$$

With a slight variation of this approach let  $\rho^*(y, s_1)$  represent the optimal risk at  $s = s_1$  subject to the restriction that fuel is not used at  $s = s_1$ . Here

$$(10.2) \quad \rho^*(y, s_1) = \inf_{y'} \{ \rho_0^*(y', s_1) + d(s_1)|y' - y| \} \quad .$$

Regarded as a function of  $y$ ,  $\rho^*$  has straight line sections with slope  $\partial \rho^* / \partial y = \pm d(s_1)$ , where it pays to use fuel. Elsewhere, it does not pay to use fuel and  $|\partial \rho^* / \partial y| \leq d(s_1)$ . Thus the optimal policy is described by an action set and a continuation or no-action set. If  $(y, s_1)$  is on the action set one moves to a point  $(\tilde{y}, s_1)$  on the boundary of the continuation set by applying fuel. Otherwise no fuel is used at this stage. It is possible to show that  $\rho^*$  is symmetric and decreasing in

$|y|$ . Hence the optimal continuation set is described by  $-\tilde{y}^*(s) < y < \tilde{y}^*(s)$ .

Let us proceed to the continuous time version of the problem for which the above characterization still applies. The boundary of the optimal no-action set is given by  $\pm\tilde{y}(s)$ . We restrict our attention to procedures which may be described by a symmetric no-action set with boundary  $\pm y_1(s)$  and let  $b(y,s)$  be the additional expected cost given  $Y(s) = y$  associated with such a procedure. We shall show that  $b(y,s)$  satisfies the following conditions

$$(10.3) \quad \frac{1}{2} b_{yy}(y,s) = b_s(y,s) \quad \text{on the no-action set}$$

$$(10.4) \quad \frac{1}{2} b_{yyy}(y,s) = b_{ys}(y,s) \quad \text{on the no-action set}$$

$$(10.5) \quad b_y(y,s) = d(s) = s^{-1} \quad \text{on that part of the boundary and action set for which } y > 0, s > 0$$

$$(10.6) \quad b_y(0,s) = 0 \quad \text{for } s > 0$$

$$(10.7) \quad b_y(y,0) = y \quad \text{for } y > 0$$

However, in Problem 4 of Section 9, we described a stopping problem whose solution uniformly minimizes  $b_y$  subject to the restrictions (10.4-10.7). Consequently the optimal expected cost for our control problem can be obtained by integrating the solution of the stopping problem, Problem 4. The optimal no-action set is the optimal continuation set of the stopping problem.

In this particular case we have been fortunate and profited from the symmetry. Otherwise we would have, for arbitrary procedures described

by no-action sets, that  $b(y,s)$  satisfies (10.3), (10.4), (10.7) and (10.5) replaced by

$$(10.5') \quad b_y(y,s) = id(s) \quad \text{on the boundary and action set}$$

The optimality condition required to determine the free boundary would be

$$(10.8) \quad \rho_{yy}(y,s) = 0 \quad \text{on the boundary}$$

This corresponds to  $(\rho_y)_y = d_y$ , and thus the derivative of the expected cost of the control problem satisfies the same (f.b.p.) as do the optimal stopping problems.

In these discussions several claims were made which require some support. First let us deal with the formulation of Problem 3\*. Suppose that incoming data used to estimate the miss distance have variance inversely proportional to the distance to target. Then the reasoning of Section 4 applied to a continuous time version indicates that at any given time, the posterior distribution of the miss distance is  $\mathcal{N}(Y(s), s)$  where  $Y(s)$ , the current estimate of the miss distance, is a Wiener process in the  $-s$  scale and  $s$ , measuring the total cumulated precision is given by

$$s^{-1} = \sigma_0^{-2} + \int_0^t \sigma^2(t_0 - t)^{-2} dt = \sigma_0^{-2} - \sigma_{t_0}^{-2-1} + \sigma^2(t_0 - t)^{-1}$$

where  $t_0$  is the total required time of flight. For simplicity let us assume that the two quantities,  $\sigma_0^{-2}$  and  $\sigma_{t_0}^{-2-1}$ , both of which are ordinarily small, cancel giving

$$s^{-1} = \sigma^2(t_0 - t)^{-1}.$$

Let us also assume that the amount of fuel required to change the miss distance by an amount  $\Delta$  is proportional to the distance to target.

Then

$$e(t) = ck'(t_0 - t) = d^{-1}(s)$$

and hence

$$(10.9) \quad d(s) = as^{-1}.$$

Since  $s = 0$  corresponds to infinite precision and time of arrival, the cost of missing is  $kY^2(0)$ . Now let  $s^* = h^2s$  and  $Y^*(s^*) = hY(s)$ . Then  $Y^*$  is a Wiener process in the  $-s^*$  scale and, in terms of  $s^*$  and  $Y^*$

$$d(s) = d^*(s^*) = ah^2s^{*-1}$$

$$kY^2(0) = kh^{-2}[Y^*(s^*)]^2.$$

Selecting  $kh^{-2} = \frac{1}{2}ah^2$  gives us a starred problem where the costs are proportional to those of Problem 3\*.

The fact that  $b$  satisfies the heat equation in the interior of the no-action set follows by the typical argument. This in turn implies that  $b_y$  satisfies the heat equation. Equations (10.6) and (10.7) follow from the symmetry and terminal cost.

To justify (10.5) one must consider behavior near the boundary. Here, one puzzling aspect of the continuous time version of our policy which was deliberately evaded must now be faced. Suppose that the

boundary  $y_1(s)$  is a well behaved function of  $s$ . Then if  $(y, s)$  is on the action set, the policy calls for bringing  $Y(s)$  to  $y_1(s)$ . Later the unadjusted  $Y(s)$  is a rather complicated function and is bound to leave the no action region "immediately" after it is brought to the boundary. How does one compute the amount of fuel that is used in the many infinitesimal departures and returns? Fortunately this can be conveniently expressed in terms of

$$(10.10) \quad M(s) = \max(0, \sup_{s_0 \geq s' \geq s} [Y_1(s') - \tilde{y}_1(s')])$$

where  $Y_1(s')$  is the original (unadjusted) Wiener process and in the neighborhood of an upper boundary point  $(y_1(s_0), s_0)$  of the no-action set, the adjusted process behaves like

$$(10.11) \quad Y(s) = Y_1(s) - M(s)$$

If  $Y(s) = y_1(s_0)$  and  $y_1(s)$  has finite slope then for small  $\delta$ ,  $\mathcal{L}(M(s_0 - \delta)) \sim \mathcal{L}(\delta^{1/2} M^*)$  where  $M^* = \sup_{0 \leq t \leq 1} W(t)$  and  $W(t)$  is a standard Wiener process. Moreover the use of a reflection principle yields  $P(M^* > a) = 2P(W(1) > a)$  for  $a > 0$  and hence

$$(10.12) \quad \mathcal{L}(M(s_0 - \delta)) \sim \mathcal{L}(\delta^{1/2} |v|) \quad \text{as } \delta \rightarrow 0$$

where  $\mathcal{L}(v) = \eta(0, 1)$ .

Now, to demonstrate (10.5) at an upper boundary point  $(y_0, s_0)$  it is easy to see that  $b_y^+(y_0, s_0) = d(s_0)$ . We shall now show that  $b_y^-(y_0, s_0) = d(s_0)$  assuming that  $b(y, s) - b(y, s - \delta) = O(\delta)$ . Between

time  $s_0$  and  $s_0 - \delta$ , the process originating from  $(y_0, s_0)$  is adjusted by a total amount  $M(s_0 - \delta)$  and at time  $s_0 - \delta$  is at  $Y(s_0 - \delta) = Y_1(s_0 - \delta) - M(s_0 - \delta)$ .

$$\begin{aligned} b(y_0, s_0) &= E[d(s_0)M(s_0 - \delta) + b(Y(s_0 - \delta), s_0 - \delta)] + O(\delta) \\ &= E[b(y_0, s_0) + b_y^- [Y_1(s_0 - \delta) - y_0 - M(s_0 - \delta)] + d(s_0)M(s_0 - \delta) + O(\delta)] \\ &= b(y_0, s_0) - [b_y^- - d(s_0)]E(M(s_0 - \delta)) + o(\delta^{1/2}) \end{aligned}$$

Since  $E(M(s_0 - \delta))$  is approximately  $(2\delta/\pi)^{1/2}$ , the desired result follows.

Finally, we demonstrate that optimality implies that  $\rho_{yy} = 0$  on the boundary. First, since  $b_y$  is constant above  $(y_0, s_0)$ ,  $\rho_{yy}^+(y_0, s_0) = 0$ . Second, since  $|p_y| \leq d(s_0)$  in the no-action set,  $\rho_y^-(y_0, s_0) \geq 0$ . The suboptimal procedure in which no action is taken from  $s_0 + \delta$  to  $s_0$ , and an optimal policy is followed thereafter has risk  $b$  where

$$\rho(y_0, s_0 + \delta) \leq b = E(\rho(y_0 + w\sqrt{\delta}, s_0))$$

$$\rho(y_0, s_0 + \delta) \leq \rho(y_0, s_0) + E \rho_y \delta + \frac{1}{2}[(\rho_{yy}^+ \delta) \int_0^\infty w^2 \phi(w) dw + (\rho_{yy}^- \delta) \int_{-\infty}^0 w^2 \phi(w) dw] + o(\delta)$$

$$\rho_s \delta = \frac{1}{2} \rho_{yy}^- \delta \leq \frac{1}{4}(\delta) [\rho_{yy}^-] + o(\delta)$$

$$\rho_{yy}^- \leq 0$$

which implies  $\rho_{yy} = 0$ .

Thus we see that  $\rho_y$  and  $\mathcal{C}$ , the partial derivative of the optimal risk and the optimal no-action set for the control problem, correspond to a solution of the free boundary problem determined by  $d(s)$ . This is the case even without the benefit of the symmetry which we used. Furthermore in the more general case where the cost of fuel per unit change of  $y$  is represented by  $d(y, s)$ , the boundary conditions for  $\rho_y$  would be  $\rho_y = d$  and  $\rho_{yy} = d_y$ .

#### 11. Summary and Remarks.

Approximating discrete time problems by continuous time problems invoking the Wiener process makes it possible to apply the analytic methods of partial differential equations to problems of sequential analysis, which are basically special examples of stopping problems, and to certain stochastic control problems. It was seen that the solution of stopping problems reduce to the solution of free boundary problems involving the heat equation. Almost arbitrary solutions of the heat equation could be used to provide bounds on the solution of stopping problems. Asymptotic expansions for the solution are obtainable by use of relatively simple classes of solutions of the heat equation.

It is difficult to state and prove rigorously "nice" theorems of the kind "the solution of the optimization problem is a solution of the f.b.p.". It would be desirable to have such theorems which invoke only conditions on the elements in the statement of the problem such as the function  $d(y,s)$ . Most proofs seem to involve conditions on the nature of the unknown solutions. This problem seems hard to avoid because the solutions of certain problems of interest have singular points where the f.b.p. condition breaks down. On the other hand the sufficiency theorems which permit one to recognize when a given candidate is a solution of the optimization problem, are much more amenable to useful statements which can be reasonably applied. Fortunately these sufficiency results are the more important ones because they are the ones invoked in applying the methods of bounding solutions of stated problems in terms of arbitrary solutions of the heat equation and solutions of related optimum problems.

The rocket control problem has a continuous time formulation similar in certain aspects to that of the stopping problem and here the derivative of the optimal expected cost also is a solution of the f.b.p.

A rocket control problem where fuel is free but only a finite amount is available, is more difficult to treat. A role analogous to that of  $\rho_y$  in the infinite fuel case is played by  $V = \rho_u + e(s)\rho_y$  where  $u$  is the amount of fuel available and  $e(s)$  is the change in  $y$  obtainable from a unit of fuel. Since  $V$  measures the rate of gain derived from using fuel,  $V \leq 0$  on the no-action set and  $V = 0$  on the action set. Bounds and expansions have been derived for the solution of this problem subject to the following conjecture. Let a taxed version of the control problem be such that at  $s = s_0$  fuel is free, but later, ( $s < s_0$ ), fuel

must be paid for and fuel remaining at  $s = 0$  must be taxed. For a situation where the original untaxed problem calls for the use of fuel at  $s = s_0$ , the taxed version also calls for the use of fuel then.

An approximation has been derived which relates the continuous time solution to discrete time solutions of stopping problems with finely spaced intervals between the permitted stopping times.

Work has been carried out on continuous time stopping problems which do not involve the heat equation. These include that of Mikhalevich [19] where the Poisson process leads to a difference differential equation and a pair of diffusion equations of Shiryaev [22] where the number of possible values of  $\mu$  are finite and past history is summarized by a few posterior probabilities rather than  $(Y(s), s)$ . Bather [4,5] has considered certain problems which involve ordinary differential equations because of the stationarity of these problems.

The history of these ideas and methods is long and complicated and the following is a bare outline of related references.

Backward induction - Dynamic programming - [2], [8].

Stopping problems (discrete time) - [14], [17], [24], [25].

Stopping problems (continuous time) - [1], [3], [4], [5], [9], [10],  
[11], [12], [13], [16], [19],  
[20], [22].

Rocket control - [6], [7], [21], [26].

Heat equation - [15].

Expository article - [23].

# REFERENCES

- [1] Anscombe, F. J. (1963). Sequential medical trials. J. Amer. Statist. Assoc., 58 365-383.
- [2] Arrow, K. J., Blackwell, D., and Girschick, M. A. (1949). Bayes and minimax solutions of sequential decision problems. Econometrica 17 213-244.
- [3] Bather, J. A. (1962). Bayes procedures for deciding the sign of a normal mean. Proc. Cambr. Phil. Soc., 58 599-620.
- [4] Bather, J. A. (1963). Control charts and the minimization of costs (with discussion). J. Royal Statist Soc. Series B 25 49-80.
- [5] Bather, J. A. (1966). A continuous time inventory model. J. Appl. Prob. 3 538-549.
- [6] Bather, J. A. and Chernoff, H. (1967). Sequential decisions in the control of a space-ship. To be published in Proc. of Fifth Berkeley Symposium.
- [7] Bather, J. A. and Chernoff, H. (1965). Sequential decisions in the control of a space-ship (finite-fuel). Stanford Tech. Rept. 14 under NSF Grant GP-3836, 1-36.
- [8] Bellman, R. (1956). Dynamic Programming, Princeton Univ. Press, Princeton, N. J.
- [9] Breakwell, J., and Chernoff, H. (1964). Sequential tests for the mean of a normal distribution II (large  $t$ ). Ann. Math. Statist. 35 162-173.
- [10] Chernoff, H. (1961). Sequential tests for the mean of a normal distribution. Proc. Fourth Berkeley Symp. Math. Statist. Prob. 1 79-91, Univ. of California Press, Berkeley.

- [11] Chernoff, H. (1965). Sequential tests for the mean of a normal distribution III (small  $t$ ). Ann. Math. Statist. 36 28-54.
- [12] Chernoff, H. (1965). Sequential tests for the mean of a normal distribution IV (discrete case). Ann. Math. Statist. 36 55-68.
- [13] Chernoff, H. and Ray, S. N. (1965). A Bayes sequential sampling inspection plan. Ann. Math. Statist. 36 1387-1407.
- [14] Chow, Y. S. and Robbins, H. E. (1967). On values associated with a stochastic sequence. To be published in Proc. of Fifth Berkeley Symp.
- [15] Doob, J. L. (1955). A probability approach to the heat equation. Trans. Amer. Math. Soc. 80 216-280.
- [16] Grigelionis, B. I. and Shiryaev, A. N. (1966). On Stefan's problem and optimal stopping rules for Markov processes. Theory of Prob. and Applications 11, 4 612-631 (English translation).
- [17] Haggstrom, G. W. (1966). Optimal stopping and experimental design. Ann. Math. Statist. 37 7-29.
- [18] Lindley, D. V. (1965). Introduction to Probability and Statistics from a Bayesian Point of View. Part II. Inference, Cambridge Univ. Press, Cambridge.
- [19] Mikhalevich, V. S. (1956). Sequential Bayes and optimal methods of statistical acceptance control. Theory of Prob. and Applications I, 4, 395-420 (English translation).
- [20] Moriguti, S. and Robbins, H. E. (1961). A Bayes test of " $p \leq 1/2$ " versus " $p > 1/2$ ". Rep. Statist. Appl. Res. Un. Japan Sci. Engrs. 9 39-60.

- [21] Orford, R. J. (1963). Optimal stochastic control systems. J. of Math. Analysis and Applications 6 419-429.
- [22] Shiryaev, A. N. (1964). On the theory of decision functions and control by a process of observation on incomplete data. Trans. III Prague Conference on Information Theory 657-681.
- [23] Shiryaev, A. N. (1965). Sequential analysis and controlled random processes (Discrete time). Kibernetika 3 1-24.
- [24] Siegmund, D. O. (1967). Some problems in the theory of optimal stopping rules. To be published in Annals of Math. Statist.
- [25] Snell, J. L. (1952). Applications of martingale system theorems. Trans. Amer. Math. Soc. 73 293-312.
- [26] Tung, F. and Striebel, C. J. (1965). A stochastic control problem and its applications. J. of Math Analysis and Applications 12 350-359.

**RELIABILITY THEORY**

by

**RICHARD E. BARLOW**

at the

**American Mathematical Society Summer Seminar**

on the

**Mathematics of the Decision Sciences**

**Stanford University**

**July - August 1967**

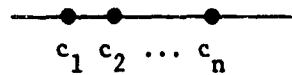
## RELIABILITY THEORY

by

Richard E. Barlow

### Closure under the Formation of Coherent Structures

Perhaps the most common structures for reliability consideration are the *series structures*



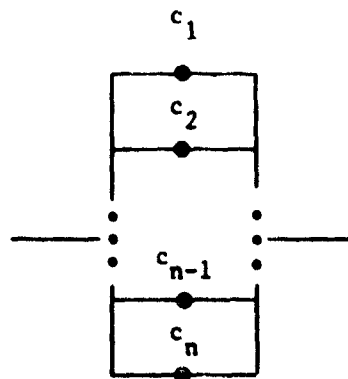
with Boolean structure function

$$\phi(\underline{x}) = x_1 x_2 \dots x_n$$

where the *indicator variable*

$$x_i = \begin{cases} 1 & \text{if the } i^{\text{th}} \text{ component, } c_i, \text{ works} \\ 0 & \text{otherwise} \end{cases}$$

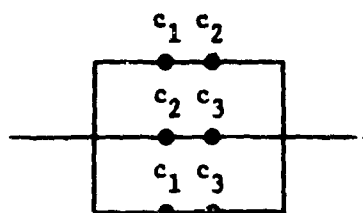
and the *parallel structures*



with Boolean structure function

$$\phi(\underline{x}) = x_1 \vee x_2 \vee \dots \vee x_n$$

where  $x \vee y = 1$  if and only if  $x = 1$  or  $y = 1$  or  $x = 1, y = 1$ . Also important are the  $k$  out of  $n$  structures which work if any  $k$  or more of the  $n$  components work. The 2 out of 3 structure, for example, can also be represented in terms of series and parallel structures if we allow replication; i.e.,



The Boolean structure function for this special structure is

$$\phi(\underline{x}) = x_1 x_2 \vee x_2 x_3 \vee x_1 x_3.$$

More generally, we have the following:

**DEFINITION** A coherent structure is a couple  $(C, \phi)$  consisting of:

- 1) a set of components  $C = \{c_1, c_2, \dots, c_n\}$ ;
- 2) a Boolean function  $\phi$  defined on vectors  $\underline{x} = (x_1, \dots, x_n)$  of binary indicator variables and satisfying

$$(i) \quad \phi(\underline{0}) = 0 \text{ and } \phi(\underline{1}) = 1;$$

$$(ii) \quad \underline{x} \leq \underline{y} \text{ (coordinatewise) implies } \phi(\underline{x}) \leq \phi(\underline{y})$$

Let  $X_i$  be a binary random variable corresponding to the  $i^{\text{th}}$  component and let

$$P\{X_i = 1\} = p_i$$

$$P[X_1 = 0] = 1 - p_1 = q_1.$$

Let  $\underline{X} = (X_1, X_2, \dots, X_n)$ ,  $\underline{p} = (p_1, p_2, \dots, p_n)$  and assume that the  $X_i$ 's are mutually independent.

**DEFINITION** The *reliability function* of the coherent structure  $(C, \phi)$  is

$$h(\underline{p}) = P[\phi(\underline{X}) = 1 \mid \underline{p}].$$

It is easy to show that  $h(\underline{p})$  is an increasing function of  $\underline{p}$  for coherent structures.

We now wish to study the *random time* at which a coherent structure fails as distinct from describing its condition at a *specified time*. To do this, let  $T_1, T_2, \dots, T_n$  denote the failure times of components. The reliability of the  $i^{\text{th}}$  component at time  $t$  is

$$\bar{F}_i(t) = P\{T_i > t\}.$$

Let

$$X_i(t) = \begin{cases} 1 & \text{if } T_i > t \\ 0 & \text{otherwise} \end{cases}$$

and let  $T$  be the time to failure of the structure. Then  $T > t$  if and only if  $\phi[\underline{X}(t)] = 1$  where  $\underline{X}(t) = (X_1(t), X_2(t), \dots, X_n(t))$ . The reliability at time  $t$  of the structure is

$$\begin{aligned} \bar{F}(t) &= 1 - F(t) = P[T > t] = P[\phi[\underline{X}(t)] = 1] \\ &= h[\bar{\underline{F}}(t)] \end{aligned}$$

where  $\bar{\underline{F}}(t) = (\bar{F}_1(t), \bar{F}_2(t), \dots, \bar{F}_n(t))$ .

Suppose we build a coherent structure from stochastically independent components whose failure times follow an exponential law; i.e.,

$$F_1(t) = 1 - \exp(-\lambda_1 t) \text{ for } \lambda_1, t > 0.$$

If the structure is a series structure then the lifetime of the structure,  $T$ , again has an exponential distribution. However, in the parallel case it is easy to verify that  $T$  is *not* exponentially distributed. What can we say in general about the properties of the distribution of  $T$ ? Birnbaum, Esary and Marshall (1966) have actually characterized this class of distributions. It is perhaps easiest to describe this class in terms of the failure rate function. Let  $F$  be a distribution on  $(0, \infty)$  with density  $f$  and

$$r(t) = \frac{f(t)}{1 - F(t)} \text{ for } t \geq 0.$$

Then, intuitively,  $r(t)dt$  is the conditional probability of failure in  $(t, t + dt)$  given survival to time  $t$ . Note that

$$\bar{F}(t) = 1 - F(t) = \exp\left[-\int_0^t r(u)du\right].$$

**DEFINITION** A distribution  $F$  such that  $F(0) = 0$  is called *IFRA* (for increasing failure rate average) if and only if

$$-\log \bar{F}(t)/t$$

is nondecreasing in  $t \geq 0$ .

If  $F$  has a density  $f$ , then it is easy to verify that  $F$  is IFRA iff

$\frac{1}{t} \int_0^t r(u)du$  is nondecreasing in  $t \geq 0$ . Note that exponential distributions are IFRA.

**THEOREM 1:** (Birnbaum, Esary and Marshall). The IFRA class of distributions is closed under the formation of coherent structures. Furthermore, the closure under coherent structures of the exponential class of distributions is dense in the IFRA class with respect to limits in distribution; i.e.,

$$\{IFRA\}^{CS} = \{IFRA\} + \{exp\}^{CS, LD}.$$

We omit the proof to this theorem. The key to the proof of this theorem is an inequality which is of independent interest.

**THEOREM 2:** (Birnbaum, Esary and Marshall). If  $h$  is the reliability function of a coherent structure and  $\psi$  is defined on  $[0, 1]$  by either

$$(i) \quad \psi(u) = -u \log u$$

$$(ii) \quad \psi(u) = -(1-u) \log (1-u) \quad (\text{the dual of (i)})$$

$$\text{or} \quad (iii) \quad \psi(u) = u(1-u)$$

then the inequality

$$(I) \quad \sum_{i=1}^n \frac{\partial h(p)}{\partial p_i} \psi(p_i) \geq \psi[h(p)]$$

holds for all  $p$  vectors.

If  $p_1 = p_2 = \dots = p_n = p$  and  $\psi(u) = u(1-u)$  then

$$\frac{dh(p)}{dp} = \sum_{i=1}^n \frac{\partial h(p)}{\partial p_i}$$

and (I) becomes

$$pq \frac{dh(p)}{dp} > h(p)(1-h(p))$$

for  $0 < p < 1$ . (The strict inequality can be shown for  $p$  in the open interval using the fact that all components are assumed essential.) If  $h(p_0) = p_0$ , then

$$h'(p_0) > \frac{h(p_0)|1-h(p_0)|}{p_0(1-p_0)} = 1;$$

i.e., at a crossing point of  $h^*(p) \equiv p$  by  $h(p)$  we see that  $h(p)$  is increasing and has slope  $> 1$  so that we have the situation in Figure 1.

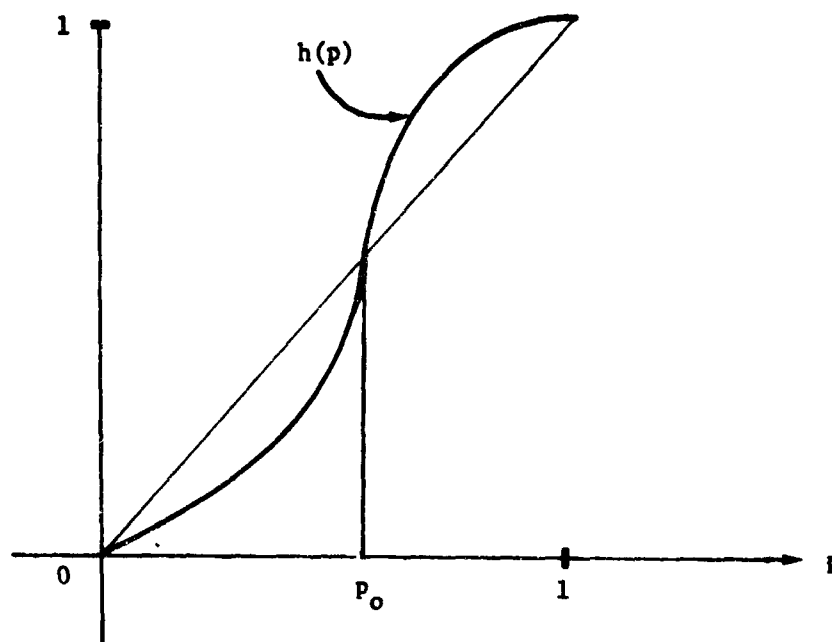


FIGURE 1

Since  $h(p)$  is increasing it can cross  $h(p) = p$  at most once and from below. The usefulness of this result follows from the fact that a redundant structure with reliability function  $h(p)$  will have higher reliability than a single component for component reliability  $p > p_0$ . This result was first discovered by Moore and Shannon (1956) for two terminal networks.

PROOF OF THEOREM 2:

The proof is by induction. For  $n = 1$ ,  $\phi(x) = x$  with  $h(p) = p$  and we have equality in (1).

Now we assume the theorem is true for  $n - 1$ . We claim (1) is true for  $h(1_n, p)$ . Either  $\phi(1_n, \underline{x})$  is coherent or  $\phi(1_n, \underline{x}) \equiv 1$ . In either case  $h(1_n, p)$  satisfies (1).

We claim (1) is true for  $h(0_n, p)$ . Either  $\phi(0_n, \underline{x})$  is coherent or  $\phi(0_n, \underline{x}) \equiv 0$ . In either case,  $h(0_n, p)$  satisfies (1).

Now  $\phi(\underline{x}) = x_n \phi(1_n, \underline{x}) + (1 - x_n) \phi(0_n, \underline{x})$  and

$$h(p) = E[\phi(X)] = p_n h(1_n, p) + (1 - p_n) h(0_n, p).$$

Also

$$\frac{\partial h(p)}{\partial p_n} = h(1_n, p) - h(0_n, p).$$

Hence

$$(2) \quad \sum_{i=1}^n \frac{\partial h(p)}{\partial p_i} \psi(p_i) = \sum_{i=1}^{n-1} \frac{\partial h(p)}{\partial p_i} \psi(p_i) + [h(1_n, p) - h(0_n, p)] \psi(p_n).$$

Now we substitute into (2) using

$$h(p) = p_n h(1_n, p) + (1 - p_n) h(0_n, p)$$

so that

$$\begin{aligned} \sum_{i=1}^n \frac{\partial h(p)}{\partial p_i} \psi(p_i) &= p_n \sum_{i=1}^{n-1} \frac{\partial h(1_n, p)}{\partial p_i} \psi(p_i) \\ &\quad + (1 - p_n) \sum_{i=1}^{n-1} \frac{\partial h(0_n, p)}{\partial p_i} \psi(p_i) \\ &\quad + [h(1_n, p) - h(0_n, p)] \psi(p_n). \end{aligned}$$

By the induction hypothesis, this is

$$\begin{aligned} &\geq p_n \psi[h(1_n, p)] + (1 - p_n) \psi[h(0_n, p)] \\ &\quad + [h(1_n, p) - h(0_n, p)] \psi(p_n). \end{aligned}$$

We will be done if we can show

$$\begin{aligned} &p_n \psi[h(1_n, p)] + (1 - p_n) \psi[h(0_n, p)] \\ &\quad + [h(1_n, p) - h(0_n, p)] \psi(p_n) \geq \psi[h(p)]. \end{aligned}$$

Let  $r = p_n$ ,  $h_1 = h(1_n, p)$  and  $h(0_n, p) = h_0$ . To show

$$\begin{aligned}
& r\psi(h_1) + (1-r)\psi(h_0) + [h_1 - h_0]\psi(r) \\
& \geq \psi[rh_1 + (1-r)h_0] ,
\end{aligned}$$

i.e., to show

$$\begin{aligned}
& \psi[rh_1 + (1-r)h_0] - \psi(h_0) \\
& \leq r\psi(h_1) - r\psi(h_0) + (h_1 - h_0)\psi(r) .
\end{aligned}$$

TO SHOW (i): Let  $\psi(u) = -u \log u$ . We claim

$$r\psi(h_1) - r\psi(h_0) + (h_1 - h_0)\psi(r) = \psi(rh_1) - \psi(rh_0) .$$

Substituting in for  $\psi(u)$  it is obvious. Hence we need only show

$$\psi[rh_1 + (1-r)h_0] - \psi[h_0] \leq \psi[rh_1] - \psi[rh_0] .$$

This is geometrically obvious from the concavity of  $\psi(u) = -u \log u$ .

TO SHOW (iii): Let  $\psi(u) = u(1-u)$ . We need only show

$$\begin{aligned}
& rh_1(1-h_1) + (1-r)h_0(1-h_0) + (h_1 - h_0)r(1-r) \\
& \geq [rh_1 + (1-r)h_0] [1 - rh_1 - (1-r)h_0]
\end{aligned}$$

or

$$-rh_1^2 - (1-r)h_0^2 + (h_1 - h_0)r(1-r) \geq -[rh_1 + (1-r)h_0]^2$$

or

$$\begin{aligned}
& r^2h_1^2 + 2r(1-r)h_0h_1 + (1-r)^2h_0^2 - rh_1^2 - (1-r)h_0^2 \\
& + (h_1 - h_0)r(1-r) \geq 0
\end{aligned}$$

or

$$-h_1^2 + 2h_0h_1 - h_0^2 + h_1 - h_0 \geq 0$$

or

$$-(h_1 - h_0)^2 + (h_1 - h_0) \geq 0$$

which is obvious.//

#### Bounds on Failure Distributions

Classes of Failure Distributions. The IFRA Failure distributions mentioned earlier are theoretically attractive because of their closure property with respect to coherent structures. They also possess an interesting graphical property which is useful in theoretical investigations. Let

$$G(x) = \begin{cases} 1 - e^{-x} & \text{for } x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Then the graph of

$$-\log[1 - F(x)] = G^{-1}F(x)$$

is starshaped with respect to the origin for  $x \geq 0$ ; i.e.,  $\frac{G^{-1}F(x)}{x} \downarrow$  in  $x \geq 0$  implies that the "upper side" of every point on the graph of  $G^{-1}F(x)$  is "visible" from the origin. Figure 2 is an illustration of a starshaped function with respect to the origin which is not convex.

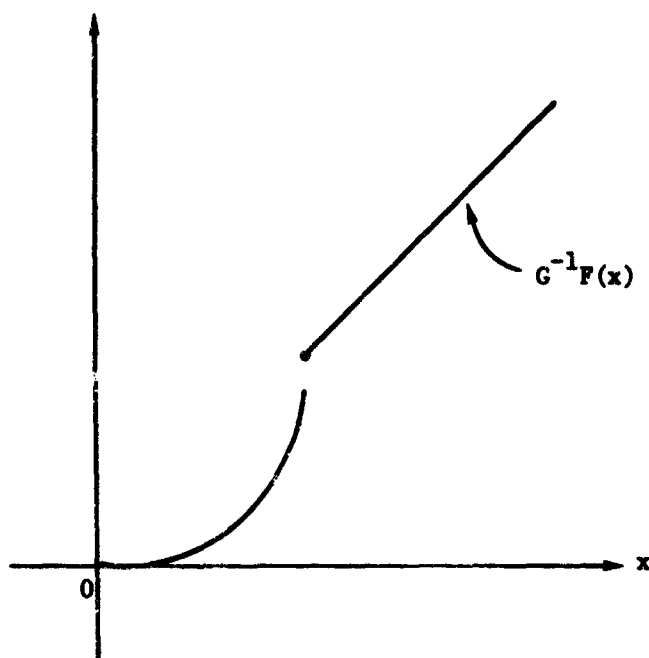


FIGURE 2.

Note that  $G^{-1}F(x)$  is convex for  $x \geq 0$  if and only if it is starshaped with respect to every point on its graph.

In replacement policy problems especially, one is often concerned with the conditional failure distribution given survival to time  $t$ ; i.e.,

$$\bar{F}_t(x) = P(X > t + x \mid X > t) = \frac{\bar{F}(t + x)}{\bar{F}(t)}.$$

It is mathematically convenient, and also intuitively plausible in some situations, that the conditional failure distribution of an assembly should also exhibit some "wearout" characteristic if the original failure distribution exhibits such a characteristic. It is, however, easy to provide examples of IFRA distributions which are not only *not* conditionally IFRA but are in fact conditionally DFRA for some  $t > 0$ . Hence we may ask the question: *What is the largest class of distributions in the IFRA class which remain IFRA upon conditioning on the left?* That is we want

$$\frac{-\log \bar{F}_t(x)}{x} = \frac{-\log \left[ \frac{\bar{F}(t+x)}{\bar{F}(t)} \right]}{x}$$

nondecreasing in  $x \geq 0$  for every  $t \geq 0$ . Clearly this will be true only if the graph of  $-\log \bar{F}(x)$  is starshaped with respect to every point on the graph. It follows that  $-\log \bar{F}(x)$  is convex for  $x \geq 0$ . If  $F$  has a density  $f$ , then the failure rate function

$$r(x) = \frac{f(x)}{1 - F(x)}$$

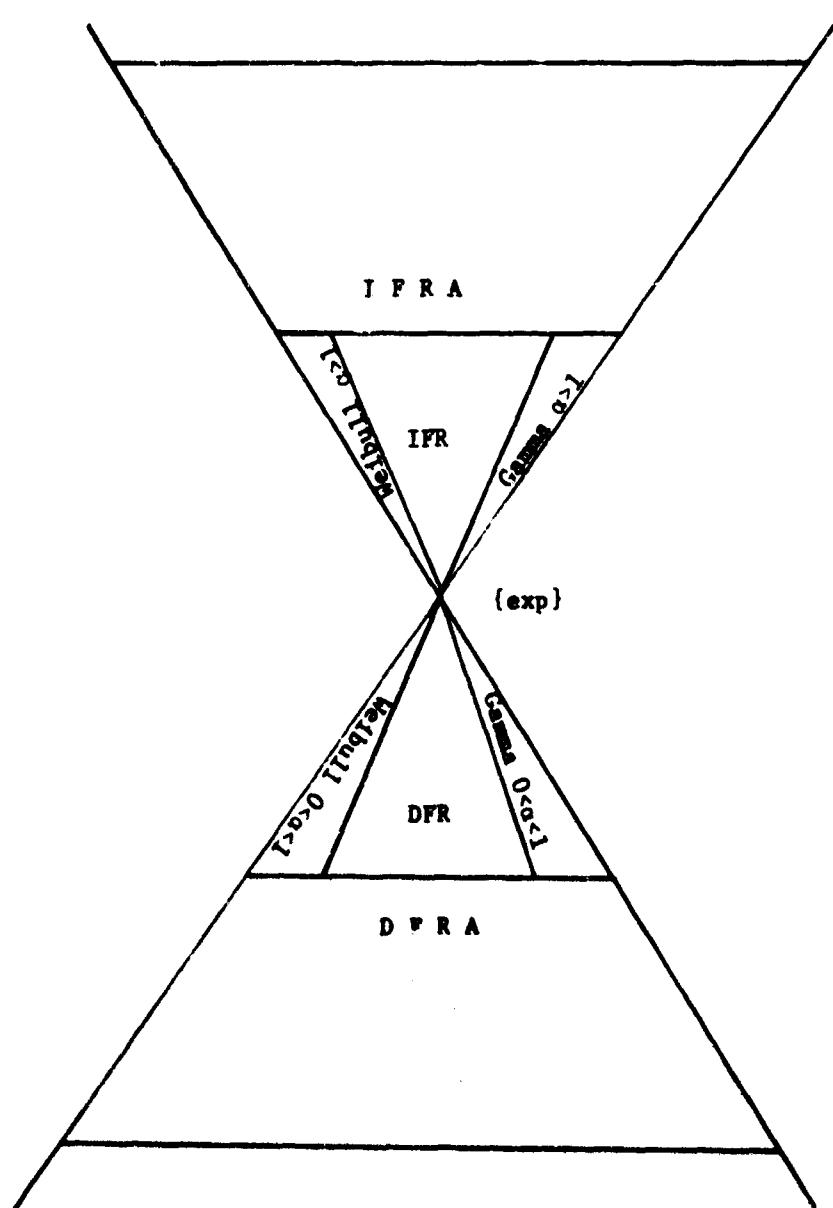
must be nondecreasing in  $x \geq 0$ . We call this class of distributions the IFR class, for increasing failure rate.

As we have seen, the exponential distribution provides the basis for an interesting hierarchy of failure distributions. The representation in Figure 3, suggested by James Esary, emphasizes the central role of the exponential distribution. Special classes noted in the figure are the Weibull class with densities

$$f(t) = \alpha \lambda t^{\alpha-1} e^{-\lambda t^\alpha} \quad \text{for } \alpha, \lambda, t \geq 0$$

and the gamma class with densities

$$f(t) = \lambda \frac{(\lambda t)^{\alpha-1}}{\Gamma(\alpha)} e^{-\lambda t} \quad \text{for } \alpha, \lambda, t \geq 0$$



**FIGURE 3: REPRESENTATION FOR FAMILIES OF LIFE DISTRIBUTIONS**

The No-Data Problem. In the aerospace and electronics industries one of a kind assemblies are common. For such assemblies no life testing results are available. All that is available in many cases is the engineers' past experience with similar components. Contractual obligations however often require a reliability statement - also warranties require a reliability assessment; e.g., it may be required that the assembly operate properly for 1000 hours with probability .99.

The mean life of an assembly is a concept with which all laymen are familiar and engineers will make statements in terms of mean life far sooner than they will make a probability statement. Hence, even in the absence of data, engineers will often estimate the mean life for a new piece of equipment based on past experience. We can translate this mean life statement into a conservative probability statement using bounds based only on intuitively reasonable assumptions. More generally, given an  $r^{\text{th}}$  moment we can state the following result (see Barlow and Marshall (1964)):

THEOREM 3: If  $F$  is IFR,  $F(0) = 0$ ,  $r \geq 1$  and  $\mu_r = \int_0^\infty x^r dF(x)$ , then

$$(1) \quad 1 - F(t) \geq \exp \left[ -t / (\lambda_r)^{\frac{1}{r}} \right] \quad \begin{array}{l} t \leq \mu_r^{\frac{1}{r}} \\ t > \mu_r^{\frac{1}{r}} \end{array}$$

$$\geq 0$$

where  $\lambda_r = \mu_r / \Gamma(r+1)$ . This inequality is sharp.

#### PROOF

We can actually prove a more general but less motivated result and the proof is easy. Suppose  $F$  and  $G$  are any two continuous distributions satisfying  $F(0) = G(0) = 0$ ,  $G^{-1}F(x)$  is convex for  $x \geq 0$  and  $\int_0^\infty x^r dF(x) = \int_0^\infty x^r dG(x) = \mu_r$  for  $r \geq 1$ . Let  $X(Y)$  have distribution  $F(G)$  and  $X^r(Y^r)$  have distribution  $F_r(G_r)$ . We claim  $G_r^{-1}F_r(x)$  is convex in  $x \geq 0$ . Note

$$G_r^{-1} F_r(x) = \left[ G^{-1} F \left( \frac{1}{x^r} \right) \right]^r$$

and assuming differentiability

$$\frac{d}{dx} G_r^{-1} F_r(x) = r \left[ \frac{G^{-1} F \left( \frac{1}{x^r} \right)}{\frac{1}{x^r}} \right]^{r-1} (G^{-1} F)' \left( \frac{1}{x^r} \right).$$

The first factor is increasing in  $x$  since  $G^{-1} F$  is starshaped and  $r \geq 1$ . The second factor is increasing in  $x$  since  $G^{-1} F$  is convex.

Now let  $X_1^r, X_2^r, \dots, X_n^r$  ( $Y_1^r, \dots, Y_n^r$ ) denote a random sample from  $F_r(G_r)$ . Then since  $G_r^{-1} F_r$  is convex

$$G_r^{-1} F_r \left[ \frac{1}{n} \sum_{i=1}^n X_i^r \right] \leq \frac{1}{n} \sum_{i=1}^n G_r^{-1} F_r(X_i^r)$$

and

$$F_r \left[ \frac{1}{n} \sum_{i=1}^n X_i^r \right] \leq G_r \left[ \frac{1}{n} \sum_{i=1}^n G_r^{-1} F_r(X_i^r) \right]$$

$$\stackrel{\text{st}}{=} G_r \left[ \frac{1}{n} \sum_{i=1}^n Y_i^r \right]$$

where  $\stackrel{\text{st}}{=}$  denotes stochastic equality. Letting  $n \rightarrow \infty$ , we have by the strong law of large numbers

$$F_r(u_r) \leq G_r(u_r)$$

or

$$F(u_r)^{\frac{1}{r}} \leq G(u_r)^{\frac{1}{r}}$$

Since  $F$  crosses  $G$  at most once and from below if at all, we have

$$F(t) \leq G(t) \quad \text{for } t \leq u_r^{\frac{1}{r}}.$$

Letting  $G(t) = 1 - \exp\left[-t/\lambda_r^{\frac{1}{r}}\right]$  for  $t \geq 0$  we easily see that  $\int_0^\infty t^r dG(t) = \nu_r$

and (1) is immediate. Since  $F$  IFR allows the possibility of a jump at the right end of its interval of support, the proof is completed using limiting arguments.

The bound for  $t > \mu_r^{\frac{1}{r}}$  is attained by the distribution degenerate at  $\mu_r^{\frac{1}{r}}$ , which is a limit of IFR distributions. //

Since it is well known that  $(\mu_r)^{\frac{1}{r}}$  is always nondecreasing in  $r > 0$  for distributions on the positive axis, we see that higher moments enable us to obtain nontrivial bounds over a greater range. To prove that  $(\mu_r)^{\frac{1}{r}}$  is nondecreasing in  $r > 0$  for distributions on the positive axis, let  $\phi(x) = x^{\frac{r'}{r}}$  where  $r \leq r'$ . Then  $\phi$  is convex for  $x \geq 0$  and

$$\begin{aligned} \phi\left[\frac{1}{n}\sum_{i=1}^n x_i^r\right] &= \left[\frac{1}{n}\sum_{i=1}^n x_i^r\right]^{\frac{r'}{r}} \\ &\leq \frac{1}{n}\sum_{i=1}^n \phi(x_i^r) = \frac{1}{n}\sum_{i=1}^n x_i^{r'} \end{aligned}$$

or

$$\left[\frac{1}{n}\sum_{i=1}^n x_i^r\right]^{\frac{1}{r}} \leq \left[\frac{1}{n}\sum_{i=1}^n x_i^{r'}\right]^{\frac{1}{r'}}.$$

where  $x_1, x_2, \dots, x_n$  is a random sample from a distribution  $F$  such that

$F(0^-) = 0$  and  $\mu_r = \int_0^\infty x^r dF(x)$ . Applying the strong law of large numbers we have

$$(\mu_r)^{\frac{1}{r}} \leq (\mu_{r'})^{\frac{1}{r'}}$$

for  $r \leq r'$ .

Unfortunately the nontrivial part of the bound in (1) is decreasing in  $r \geq 1$ .

This follows from the fact that for IFRA distributions  $\lambda_r^{\frac{1}{r}} = \left[\frac{\mu_r}{r(r+1)}\right]^{\frac{1}{r}}$  is decreasing in  $r > 0$ , or equivalently,  $-\log \lambda_r$  is starshaped for  $r > 0$ . It is interesting that for IFRA and IFR distributions, the geometrical properties of

$\bar{F}(x)$  are inherited by the "normalized" moments  $\lambda_r = \mu_r / \Gamma(r+1)$ ; i.e., if  $-\log \bar{F}(x)$  is starshaped (convex) in  $x \geq 0$ , then  $-\log \lambda_r$  is starshaped (convex) in  $r > 0$ .

**THEOREM 4:** If  $F$  is IFRA and  $\mu_r = \int_0^\infty x^r dF(x)$  then  $\lambda_r^{\frac{1}{r}} = \left( \frac{\mu_r}{\Gamma(r+1)} \right)^{\frac{1}{r}}$  is nonincreasing in  $r > 0$ .

PROOF

$$\begin{aligned} \text{Let } s \leq t \text{ and note } \mu_s &= s \int_0^\infty x^{s-1} \bar{F}(x) dx \\ &= s \int_0^\infty x^{s-1} e^{-x/\lambda_s^{\frac{1}{s}}} dx. \end{aligned}$$

Since  $F$  is IFRA,  $x^{s-1} \bar{F}(x)$  crosses  $x^{s-1} e^{-x/\lambda_s^{\frac{1}{s}}}$  exactly once and from above, say at  $x_0$ . Hence if  $\phi$  is an increasing function

$$\begin{aligned} &\int_0^\infty \phi(x) x^{s-1} \bar{F}(x) dx - \int_0^\infty \phi(x) x^{s-1} e^{-x/\lambda_s^{\frac{1}{s}}} dx \\ &= \int_0^\infty [\phi(x) - \phi(x_0)] \left[ x^{s-1} \bar{F}(x) - x^{s-1} e^{-x/\lambda_s^{\frac{1}{s}}} \right] dx \leq 0. \end{aligned}$$

Let  $\phi(x) = x^{t-s}$ . Then

$$\begin{aligned} \frac{s}{t} \mu_t &= \int_0^\infty x^{t-s} \left[ s x^{s-1} \bar{F}(x) \right] dx \leq \int_0^\infty x^{t-s} \left[ s x^{s-1} e^{-x/\lambda_s^{\frac{1}{s}}} \right] dx \\ &= \frac{s}{t} \Gamma(t+1) \left( \lambda_s^{\frac{1}{s}} \right)^t \end{aligned}$$

or  $(\lambda_t)^{\frac{1}{t}} \leq (\lambda_s)^{\frac{1}{s}}$  which was to be proved. //

Additional probability bounds may be found in Barlow and Marshall (1964), (1965) and (1966).

# REFERENCES

- Barlow, R. E. and A. W. Marshall (1964). Bounds for Distributions with Monotone Hazard Rates, I. Annals of Math. Statist., 35, 1234-57.
- Barlow, R. E. and A. W. Marshall (1965). Tables of Bounds for Distributions with Monotone Hazard Rate. J. Amer. Statist. Assoc., 60, 872-890.
- Barlow, R. E. and F. Proschan (1965). Mathematical Theory of Reliability. J. Wiley and Sons, New York.
- Barlow, R. E. and A. W. Marshall (1966). Bounds on Interval Probabilities for Restricted Families of Distributions. Proc. of the Fifth Berkeley Symposium.
- Birnbaum, Z. W. and J. D. Esary (1965). Modules of Coherent Binary Systems. J. Soc. Indust. Appl. Math., 13, 444-462.
- Birnbaum, Z. W., J. D. Esary and A. W. Marshall (1966). A Stochastic Characterization of Wear-Out for Components and Systems. Ann. Math. Statist., 37, 816-825.
- Moore, E. F. and C. E. Shannon (1956). Reliable Circuits Using Less Reliable Relays. J. of the Franklin Institute, 262, Part I, 191-208.

**MARKOVIAN DECISION PROCESSES**

by

**CYRUS DERMAN**

at the

**American Mathematical Society Summer Seminar**

on the

**Mathematics of the Decision Sciences**

**Stanford University**

**July - August 1967**

# MARKOVIAN DECISION PROCESSES - AVERAGE COST CRITERION <sup>1</sup>

by

Cyrus Derman  
Columbia University

## 1. Introduction

We are concerned with the optimal control of certain types of dynamic systems. We assume such a system is observed periodically at times  $t=0,1,\dots$ . After each observation the system is classified into one of a possible number of states. Let  $I$  denote the space of possible states.  $I$  will be assumed to be either finite or denumerable. After each classification one of a possible number of decisions is implemented. Let  $K_i$  denote the number of possible decisions when the system is in state  $i$ ,  $i \in I$ . The sequence of implemented decisions interacts with the chance environment to effect the evolution of the system.

<sup>1</sup> This research was supported by the Army, Navy, Air Force and NASA under a contract administered by the Office of Naval Research; Contract Nonr 266(55) - NR-042-099. Reproduction in whole or in part is permitted for any purpose of the United States Government.

More specifically let  $\{Y_t\}$ ,  $t=0,1,\dots$  denote the successive observed states of the system. Let  $\{\Delta_t\}$ ,  $t=0,1,\dots$  denote the successive decisions, ~~implemented after observing  $Y_t$~~ . Assume that when  $Y_t = i$  and  $\Delta_t = k$  a known cost  $w_{ik}$  is incurred. The numbers  $\{w_{ik}\}$  may be expected costs rather than actual costs. In such a case we assume a distribution of costs dependent upon the state  $i$  and decision  $k$  from which the expected cost can be computed.

A rule or policy  $R$  for controlling the system is a set of functions  $\{D_k(h_{t-1}, Y_t)\}$ ,  $t=0,\dots$ , where  $h_t = \{Y_0, \Delta_0, \dots, Y_t, \Delta_t\}$ ,  $D_k(\cdot) \geq 0$ , and  $\sum_k D_k(\cdot) = 1$ ;  $D_k(h_{t-1}, Y_t)$  is to be interpreted as the probability of implementing decision  $k$  at time  $t$  given the "history"  $h_{t-1}$  and the "present state"  $Y_t$ . Thus a rule specifies, at each point in time, a chance mechanism to be used in deciding which action to take. The rule is only permitted to depend on the history of states and decisions.

Given a rule  $R$  and a probability distribution over the initial state  $Y_0$ , we assume the sequence  $\{Y_t, \Delta_t\}$  is a stochastic process defined over the joint space of  $I$  and the possible decisions. Throughout we shall assume that  $P(Y_0 = i)$ ,  $i \in I$ , is known. Moreover, we assume that there

are known transition probabilities  $\{ q_{ij}(k) \}$  such that for every  $i, j$  and  $k$

$$q_{ij}(k) = P \{ Y_{t+1} = j \mid h_{t-1}, Y_t = i, \Delta_t = k \}$$

independent of  $t$  and  $h_{t-1}$ . In words, when, at any time  $t$ , the state  $i$  is observed and decision  $k$  is made, then  $q_{ij}(k)$  denotes the probability that the system will be observed in state  $j$  at time  $t + 1$ . Under this latter assumption we refer to the process  $\{ Y_t, \Delta_t \}$  as a Markovian

Decision Process. It shall be emphasized, however, that  $\{ Y_t, \Delta_t \}$  is not necessarily a Markov process. For when  $R$  is such that  $D_k(h_{t-1}, Y_t)$  is a function of  $h_{t-1}$ , the process  $\{ Y_t, \Delta_t \}$  will not be Markovian. If, however,  $D_k(h_{t-1}, Y_t)$  is a function of  $Y_t$  and  $t$  for every  $t=0,1,\dots$ , then  $\{ Y_t \}$  is a Markov process. And, if  $D_k(h_{t-1}, Y_t)$  is only a function of  $Y_t$  for every  $t$  then  $\{ Y_t \}$  is a Markov chain with stationary transition probabilities.

Let  $W_t$ ,  $t = 0, 1, \dots$  be defined as follows:

$$W_t = w_{ik} \text{ if } Y_t = i, \Delta_t = k, \quad k = 1, \dots, K_i, i \in I.$$

Given a policy  $R$  and an initial state  $Y_0 = i$ , then the sequence  $\{W_t\}$ ,  $t = 0, 1, \dots$ , is a stochastic process. We can speak of the expected cost at time  $t$  as

$$E_R W_t = \sum_j \sum_k w_{jk} P_R \{ Y_t = j, \Delta_t = k \mid Y_0 = i \}$$

where  $E_R$  and  $P_R$  denotes the expectation and probability under the policy  $R$ . We, of course, assume that the costs  $\{w_{ik}\}$  are such that  $E_R W_t$  exists.

$$\text{Let} \quad \varphi_{R,T}(i) = \frac{1}{T+1} \sum_{t=0}^T E_R W_t ;$$

i.e.  $\varphi_{R,T}(i)$  is the expected average cost incurred by the system up to time  $T$  given  $Y_0 = i$  and  $R$  is the policy controlling the system.

$$\text{Let} \quad \varphi_R(i) = \lim_{T \rightarrow \infty} \inf_{R,T} \varphi_{R,T}(i) .$$

The problem under consideration in this exposition is that of finding  $R$  to minimise  $\varphi_R(i)$ .

For what follows it is convenient to consider three classes of policies. The first is the class of all

policies of the form described. That is, all policies which, at each point in time, use past states and decisions as a basis for making a decision. We let  $G$  denote this class. The second class is the class which uses, at each point in time, the state of the system at that instant as a basis for making a decision. We shall refer to this class as the class of stationary Markovian policies and denote this class by  $G'$ . The third class is the sub class of  $G'$  in which the policies are not of a random character. We denote this class by  $G''$ . A policy  $R \in G''$  may be thought of as a function defined over the states with range in the set of possible decisions; to each state there corresponds a unique decision. We refer to  $G''$  as the class of deterministic stationary Markovian policies.

We shall divide the following discussion into two parts. One for the case where  $I$  is finite; the other, for where  $I$  is denumerable. In going from the finite to the denumerably infinite, mathematical questions arise which are not yet settled.

## 2. Finite Number of States.

The fundamental fact concerning the problem at hand for this finite state case can be summarized as

**Theorem I:** If  $K_i < \infty$ ,  $i \in I$ , and  $I$  is finite, then there exists a policy  $R \in G''$  which minimizes  $\phi_R(i)$ ,  $i \in I$ .

The above theorem has been more or less, proved, by several Authors (See [1], [2], [3], [4]). Each proof relies on what we refer to as

Theorem 2. If  $K_1 < \infty$ ,  $i \in I$ , and  $I$  is finite, then there exists a policy  $R_\alpha \in C^n$  which minimizes

$$v_R(i, \alpha) = \sum_{t=0}^{\infty} \alpha^t E_R w_t, \quad i \in I$$

where  $\alpha$  is a given number between zero and one.

$v_R(i, \alpha)$  is often of economic relevance. It is referred to as the expected discounted (with discount factor  $\alpha$ ) cost criterion.

Theorem 2 is usually taken as self-evident. However, for proof see [5] and [6]. On letting  $\alpha \rightarrow 1$  and using an appropriate Tauberian theorem (e.g.  $\lim_{\alpha \rightarrow 1} (1-\alpha) v_R(i, \alpha) = \varphi_R(i)$  when  $R$  is such that  $\varphi_R(i) = \lim_{T \rightarrow \infty} \varphi_{R,T}^{(i)}$ ) theorem 1 can be established.

Theorem 1 carries with it two advantages. The class  $C''$  contains only a finite number of policies (even though the number may be astronomically large) and under any  $R \in C''$ ,  $\{Y_t\}$  is a Markov chain with stationary transition probabilities. As a result, finite algorithms for minimizing  $\varphi_R(i)$  over  $R \in C''$  are obtainable.

Two methods for obtaining the optimal  $R \in C''$  have been advanced. One method involves linear programming. (See [7], [3]). The other is a derivative of dynamic programming (See [8], [2]). We indicate the latter method first.

Let  $R \in C''$  be an arbitrary policy. It can be shown that there exists a unique set of numbers

$$\{g_i^R, v_i^R\}, i \in I \text{ satisfying}$$

$$(1) \quad g_i^R + v_i^R = w_{iR} + \sum_j q_{ij}(R) v_j^R, i \in I,$$

and

$$(2) \quad \sum_j p_{ij}^R v_j^R = 0, i \in I,$$

where  $w_{iR}$  and  $q_{ij}(R)$

denote the cost  $w_{ik}$  and transition probability  $q_{ij}(k)$  involved at state  $i$  under policy  $R$ ;  $\pi_{ij}^R$  is the limiting (as  $T \rightarrow \infty$ ) expected proportion of time that the system is in state  $j$  given it is initially in state  $i$ . The theory of Markov chains indicates how the limiting values  $\{\pi_{ij}^R\}$  can be obtained.

One can also, see that  $g_i^R = \sum_{j \in I} \pi_{ij}^R w_{jR} = \varphi_R(i)$ .

For each  $i \in I$ , let  $E_i$  denote the set of decisions  $k$  for which either (a)

$$\sum_j q_{ij}(k) g_j^R < g_i^R$$

or (b)

$$\sum_j q_{ij}(k) g_j^R = g_i^R$$

and

$$w_{ik} + \sum_j q_{ij}(k) v_j^R < g_i^R + v_i.$$

Define a policy  $R'$  as follows: For at least one  $i$  such that  $E_i$  is non-empty prescribe a decision in  $E_i$  when in state  $i$ . For all states  $i$  where  $E_i$  is empty or where one

does not prescribe a decision in  $E_i$ , make the decision dictated by policy  $R$ . We shall call the mapping from  $R$  to  $R'$  a policy iteration.

It can be shown that  $\varphi_{R'}(i) \leq \varphi_R(i)$ ,  $i \in I$ . The equality may hold because of the possible presence of transient states in the Markov chain associated with policy  $R'$ . Thus the policy improvement procedure starts with an arbitrary policy  $R_0$  and carries out successive policy iterations until no more can be made. Since there are only a finite number of policies in  $C$  this stage must be reached. (Actually, this is not obvious because of the possibility that  $\varphi_{R'}(i) = \varphi_R(i)$ ,  $i \in I$ . That is, an argument must be made to show that cycles of policies will not occur.) At the termination of the successive policy iterations we must have, if  $R^*$  is the terminal policy,

$$(3) \quad g_i^{R^*} + v_i^{R^*} = \min_k \left\{ w_{ik} + \sum_j p_{ij}(k) v_j^{R^*} \right\}, \quad i \in I$$

and

$$(4) \quad \sum_j p_{ij} g_j^{R^*} \geq g_i^{R^*}, \quad i \in I.$$

From these <sup>system</sup> equations it can be shown that  $R^*$  is optimal. Thus, the sequence of policy iterations terminates at an optimal policy within a finite number of iterations.

The policy improvement procedure simplifies under the assumption

(A)  $I$  contains at most one ergodic class of states for every  $R \in C$ .

Under (A),  $g_i^R = c_i^R$  independent of the initial state  $i$ , part (a) of the definition of  $\mathcal{V}_i$  is unnecessary, and (4) of the terminal ~~equations~~<sup>by Term</sup> will always hold.

Let  $\varphi_R^{(m)}(i)$ ,  $m=1, \dots, M$  be defined in a manner similar to  $\varphi_R(i)$  except that  $\varphi_R^{(m)}(i)$  is defined with respect to costs  $\{w_{ik}^{(m)}\}$ ,  $m=1, \dots, M$ . Although the policy improvement procedure is powerful enough to obtain an optimal solution to the problem under discussion, under general conditions (provided the number of states is not too large), it does not provide an algorithm for solving the more complicated problem:

$$\text{Minimize } \varphi_R(i)$$

subject to

$$\varphi_R^{(m)}(i) \geq b_m, \quad m = 1, \dots, M$$

where  $b_m$ ,  $m = 1, \dots, M$  are given constants.

When (A) holds the method of linear programming is effective for obtaining an optimal policy. Under (A)

$$\pi_{ij}^R = \pi_j^R, \text{ independent of } i, \varphi_R(i) = \sum_j \sum_k \pi_j^R D_{jk}^R w_{jk},$$

$$\varphi_R^{(m)}(i) = \sum_j \sum_k \pi_j^R D_{jk}^R w_{jk}^{(m)}, m = 1, \dots, M,$$

where  $D_{ik}^R$  denotes the probability of making decision  $k$  when in state  $i$ ; the  $\{D_{ik}^R\}$  define a rule  $R \in C'$ .

The  $\{\pi_j^R\}$  must satisfy the steady state equations:

$$\pi_j^R \geq 0$$

$$\pi_j^R - \sum_i \sum_k \pi_i^R D_{ik}^R \varphi_{ik}(k) = 0$$

$$\sum_j \pi_j^R = 1$$

If one makes the transformation  $x_{ik} = \pi_i^R D_{ik}^R$ , one gets the linear programming problem

$$\text{Min} \quad \sum_i \sum_k x_{ik} w_{ik}$$

subject to  $x_{ik} \geq 0 \quad k=1, \dots, K_i, \quad i \in I$

$$\sum_k x_{jk} - \sum_i \sum_k x_{ik} q_{ij}(k) = 0, \quad i \in I$$

$$\sum_j \sum_k x_{jk} = 1$$

$$\sum_i \sum_k x_{ik} w_{ik}^{(m)} \geq b_m, \quad m = 1, \dots, M$$

If  $\{x_{jk}^*\}$  is a solution to the above problem then the optimal policy  $R^* \in C'$  is defined by setting

$$D_{ik}^{R^*} = \frac{x_{ik}^*}{\sum_k x_{ik}^*} \quad \text{if} \quad \sum_k x_{ik}^* > 0,$$

and arbitrary if  $\sum_k x_{ik}^* = 0$ .

If  $M = 0$ , i.e. no additional constraints are imposed, the method is an alternative to the policy iteration procedure under condition (A).

The question of how to solve the problem with additional constraints without assuming (A) remains open. In general, an optimal policy need not exist in  $C'$ . One can assert, however, (See [9]) that under (A) an optimal policy will <sup>always</sup> exist in  $C'$ .

It should be pointed out that the solution to the problem of minimizing  $\phi_R(i)$  need not be unique. The question arises as to whether some solutions might not be better than others. That is, other criteria, not explicitly put into the problem, may, in part, be relevant. For example, it has been shown (see [2]), that there exists a policy  $R^* \in C''$  such that  $\phi_R(i, \alpha)$  is minimized for all  $\alpha$  and near enough to 1. When this is the case  $R^*$  also minimizes  $\phi_R(i)$ . No computational procedure has yet been given to find such a policy. However, a procedure (see [10]) has been given for finding a policy  $R^{**}$  having the property that

$$\lim_{\alpha \rightarrow 1} [\phi_{R^*}(i, \alpha) - \phi_{R^{**}}(i, \alpha)] = 0.$$

Needless to say,  $R^{**}$  also minimizes  $\phi_R(i)$ .

### 3. Denumerable State Case

Let us turn from the finite case to the case where  $I$  is denumerably infinite and ask ourselves whether the basic facts as put forth in theorems 1 and 2 remains the same.

The following modification of theorem 2 can be shown (see [5] and [6]).

**Theorem 2'.** If  $K_i < \infty$ ,  $i \in I$ , and  $\{w_{ik}\}$  bounded, then there exists a policy  $R_\alpha \in C''$  which minimizes  $\psi_R(i, \alpha)$ ,  $i \in I$ .

If  $K_i = \infty$  it can be easily shown that theorem 2' need not hold. If the  $\{w_{ik}\}$  are not bounded it can also be shown that the result does not hold in general (see [6]). Thus, we might ask if the conclusions of theorem 1 hold under the conditions of theorem 2'.

An example [11] has been given showing that the conditions of theorem 2' do not guarantee the conclusions of theorem 1. Moreover, the example shows that an optimal policy  $R$  for minimizing  $\phi_R(i)$  may not exist. This counter-example also implies that there may not be a policy  $R \in C''$  which minimizes  $\psi_R(i, \alpha)$  for all  $\alpha$  near enough to 1, as in the case when  $I$  is finite. For if such were the case, it would be possible to show that  $\phi_R(i)$  could always be optimized.

A more surprising counter-example (see [12] ) shows that an optimal policy may not exist in  $C''$ ; but it may in  $C'$ . Thus, if restricted to stationary policies, a policy involving randomization may prove to be more effective than one that is deterministic. The counter-example exploits the fact that denumerable state Markov chains may have recurrent null states--a property denied finite state chains.

An even more surprising counter-example (see [13] ) shows that  $\varphi_R(i)$  may be minimized by a policy in  $C-C'$ , whereas no optimal policy exists in  $C'$ . Thus, in order to obtain an optimal policy one may find it necessary to go beyond the class of stationary Markovian policies. This fact seems to run counter to one's intuition regarding the problem under discussion. The literature has numerous remarks asserting the reasonableness of assuming that an optimal policy is stationary.

What develops as an interesting mathematical question is that of determining the weakest conditions under which it can be asserted that a policy  $R \in C''$  is optimal. In [12] and [14] the following was proved.

**Theorem 3.** Suppose  $\{w_{ik}\}$  are bounded. If there exists bounded numbers  $g, \{v_j\}, j \in I,$  satisfying

$$(5) \quad g + v_i = \min_k \left\{ w_{ik} + \sum_j q_{ij} v_j \right\}, i \in I.$$

then there exists a policy  $R^* \in C''$  which is optimal. The policy  $R^*$  is: implement decision  $k = k_i$  which minimizes the right side of (5) for each  $i \in I$ . Also,

$$g = \varphi_{R^*}(i), \quad i \in I.$$

Note that (5) is related to (1) with  $g_i \neq g$ . Actually, theorem 3 can be generalized to the case where the  $g_i$ 's are not all equal and (1) and (2) replace (5).

Conditions implying the hypothesis of theorem 3 can be given. The approach is to define a policy improvement procedure similar to the one discussed in the finite state case and show that in the limit one gets a policy  $R \in C''$  satisfying (5).

We define a policy iteration as follows:

Let  $R \in C''$  be given. Assume a solution  $g^R, \{v_j^R\}, j \in I$  to the system of equations

$$(6) \quad g^R + v_i^R = w_{iR} + \sum_j q_{ij}(R) v_j^R, \quad i \in I$$

exists. (The system (6) is treated in [15].) Define  $R'$  by choosing, for each  $i \in I$ , that decision  $k = k_i$  which minimizes

$$w_{ik} + \sum_j q_{ij}(k) v_j^R.$$

The transformation from  $R$  to  $R'$  is the policy iteration. Note that the policy iteration, here, is defined more stringently than for the finite state case. By a policy

improvement procedure we mean starting with an arbitrary policy  $R_0 \in C''$ , letting  $R_{n+1}$  denote the policy obtained by a policy iteration on  $R_n$ ,  $n = 0, 1, \dots$ . If, for any  $n$ ,  $R_{n+1} = R_n$ , then equations (5) are satisfied and  $R_n$  is optimal. Otherwise, one may or may not obtain an optimal policy as  $n \rightarrow \infty$ . It is of interest to provide conditions under which  $\{R_n\}$  converges to an optimal policy.

Let us list the following conditions.

(B) For every  $R \in C''$ , the associated Markov chain is irreducible and positive recurrent.

(C) For every  $R \in C''$  there exists a bounded solution

$$g^R, \{v_j^R\}, j \in I \text{ to (6). The solutions are}$$

uniformly bounded over  $R \in C''$ .

(D) For every  $i \in I$ ,  $\inf_{R \in C''} \pi_i^R > 0$ .

We can assert (see [12])

Theorem 4. If  $K_i < \infty$ ,  $i \in I$ ,  $\{w_{ik}\}$  are bounded, and

(B), (C), and (D) hold, then the policy improvement procedure converges to an optimal policy  $R^* \in C''$ .

The proof of theorem 4 involves showing that the limiting policy does yield a solution to (5). Under weaker conditions, i.e. (B) and (C), it can be shown

that a solution to (5) exists. Therefore, an optimal policy is in  $C''$ . However, it is not clear that a policy improvement procedure will converge to an optimal policy.

Conditions are given in [15] guaranteeing (C); slightly weaker conditions may be given in [13]. A better approach to the existence question would, in all likelihood, avoid the equations (5) and (6).

# REFERENCES

- [1] Gillette, Dean (1957). Stochastic games with zero stop probabilities. Ann. Math. Studies, 39, Vol. III 179-187.
- [2] Blackwell, David (1962), Discrete dynamic programming. Ann. Math. Statist. 33, 719-726.
- [3] Derman, Cyrus (1962). On sequential decisions and Markov chains. Management Science 9, 16-24.
- [4] Viskov, O.V. and Surjaev, A.M. (1964), On controls leading to optimal stationary states. Selected Translations in Mathematical Statistics and Probability, Vol. 6 71-83.
- [5] Blackwell, David (1965). Discounted dynamic programming. Ann. Math. Statist. 36, 226-235.
- [6] Derman, Cyrus (1965). Markovian Sequential Control Processes--Denumerable State Space. J. Math. Analytical Application, 10, 295-302.
- [7] Manne, Alan S. (1960). Linear Programming and Sequential Decisions. Management Science Vol. 6, No.3.
- [8] Howard, Ronald A. (1960). Dynamic Programming and Markov Processes. John Wiley, New York.
- [9] Derman, Cyrus (1963). Stable sequential control rules and Markov chains. J. Math. Analytical Application, 6, 257-265.
- [10] Veinott, Arthur F., Jr. (1966) On finding optimal policies in discrete dynamic programming with no discounting. Ann. Math. Statist. 37, 1284-1294.
- [11] Maitra, Ashok (1964). Dynamic Programming for Countable State Systems. Doctoral thesis, University of California, Berkeley.
- [12] Derman, Cyrus (1966). Denumerable State Markovian Decision Processes - Average Cost Criterion. Ann. Math. Statist. 37, 1545-1553.

- [13] Ross, Sheldon M. (1967). Non-Discounted Denumerable Markovian Decision Models. Tech. Report No. 94 Dept. of Statistics, Stanford University.
- [14] Derman, Cyrus and Lieberman, G. J. (1967). A Markovian Decision Model for a Joint Replacement and Stocking Problem. Management Science. Vol. 13, No. 9, 609-617.
- [15] Derman, Cyrus and Veinott, Arthur F. Jr. (1967). A solution to a countable system of equalities arising in Markovian decision processes. Ann. Math. Statist. 38, 582-584.

**Lectures on  
LEARNING THEORY**

**by  
M. FRANK NORMAN**

**at the  
American Mathematical Society Summer Seminar  
on the  
Mathematics of the Decision Sciences  
Stanford University  
July - August 1967**

**SOME CONVERGENCE THEOREMS FOR STOCHASTIC  
LEARNING MODELS WITH DISTANCE DIMINISHING OPERATORS<sup>1</sup>**

**M. Frank Norman**

**University of Pennsylvania, Philadelphia, Pennsylvania**

**To be published in the Journal of Mathematical Psychology, fall, 1967**

## ABSTRACT

A broad mathematical framework is considered that includes stochastic learning models with distance diminishing operators for experiments with finite numbers of responses and simple contingent reinforcement. Convergence theorems are presented that subsume most previous results about such models, and extend them in a variety of ways. These theorems permit, for example, the first treatment of the asymptotic behavior of the general linear model with experimenter-subject controlled events and no absorbing barriers. Some new results are also given for certain two-process discrimination learning models and for models with finite state spaces.

## 1. INTRODUCTION

Suppose that a subject is repeatedly exposed to an experimental situation in which various responses are possible, and suppose that each such exposure<sup>or trial</sup> can alter the subject's response tendencies in the situation. It is assumed that the subject's response tendencies on trial  $n$

are determined by his state  $S_n$  at that time. The set of possible states is denoted  $S$  and called the state space. The effect of the  $n^{\text{th}}$  trial is represented by the occurrence of a certain event  $E_n$ . The set of possible events is denoted  $E$  and referred to as the event space. The quantities  $S_n$  and  $E_n$  are to be considered random variables. The corresponding small letters  $s_n$  and  $e_n$  are used to indicate particular values of these variables, and, in general,  $s$  and  $e$  denote elements of the state and event spaces, respectively.

To represent the fact that the occurrence of an event effects a change of state, with each event  $e$  is associated a mapping  $f_e(\cdot)$  of  $S$  into  $S$  such that, if  $E_n = e$  and  $S_n = s$ , then  $S_{n+1} = f_e(s)$ . Thus

$$H1 \quad S_{n+1} = f_{E_n}(S_n)$$

for  $n \geq 1$ . The function  $f_e(\cdot)$  will be called the operator for the event  $e$  or simply an event operator. Throughout the paper it is

Norman

4.

assumed that

H2  $E$  is a finite set.

It is further supposed that the learning situation is memory-less and temporally homogeneous, in the sense that the probabilities of the various possible events on trial  $n$  depend only on the state on trial  $n$ , and not on earlier states or events, or on the trial number. That is, there is a real valued function  $\phi_{\cdot}(\cdot)$  on  $E \times S$  such that

H3  $P_s(E_1 = e_1) = \phi_{e_1}(s)$ , and

$$P_s(E_{n+1} = e_{n+1} | E_j = e_j, 1 \leq j \leq n) = \phi_{e_{n+1}}(f_{e_1 \dots e_n}(s))$$

for  $n \geq 1$ , where

$$f_{e_1 \dots e_n}(s) = f_{e_n}(f_{e_{n-1}}(\dots(f_{e_1}(s)))). \quad (1.1)$$

Throughout the paper state subscripts on probabilities and expectations are initial states, that is, values of  $S_1$ .

Two examples will be discussed in Section 3: a linear model for ordinary two-choice learning, and a two-stage linear discrimination learning model. In the first linear model, the state is the probability of one of the responses, so  $S = [0, 1]$ . In the linear discrimination learning model the state is a pair of probabilities that determine, respectively, the "response"

probabilities at the two stages. Thus,  $S = [0, 1] \times [0, 1]$ . In these examples each event involves the subject's overt response (suitably coded), the observable outcome of that response (i.e., the experimenter's response), and, sometimes, a hypothetical occurrence that is not directly observable (e.g., the state of attention on a trial). The force of assumption H3 for the experimenter is to limit reinforcement schedules to those in which the outcome probabilities depend only on the immediately preceding response, that is, to simple contingent schedules.

The research reported in this paper is directed toward understanding the asymptotic behavior of the stochastic processes  $\{S_n\}$  and  $\{E_n\}$  for a class of models with distance diminishing event operators to be defined below by imposing additional restrictions on the functions  $f$  and  $\phi$ . This class generalizes the familiar linear models, and the latter provide much of the motivation for the axioms for the former.

To discuss "distance diminishing" event operators,

it is necessary to assume that  $S$  is a metric space with respect to some metric  $d$ . A formulation

in terms of Euclidean space and root-sum-square distance would yield sufficient generality to cover the linear models of Section 3. Such a formulation would, however, restrict generality without any redeeming simplification.

Moreover, a treatment in terms of general metric spaces highlights those aspects that are crucial to the theory. For these reasons only it is assumed that

H4  $(S, d)$  is a metric space.

The reader who prefers a Euclidean setting can easily specialize most of what follows to suit his preferences. The next assumption is suggested by the linear examples of Section 3:

H5  $(S, d)$  is compact.

The remaining hypotheses are most easily stated in terms of the following notations. If  $\psi$  and  $g$  are mappings of  $S$  into the real numbers and into  $S$ , respectively, their maximum "difference quotients"  $m(\psi)$  and  $\mu(g)$  are defined by

$$m(\psi) = \sup_{s \neq s'} \frac{|\psi(s) - \psi(s')|}{d(s, s')}, \text{ and} \quad (1.2)$$

$$\mu(g) = \sup_{s \neq s'} \frac{d(g(s), g(s'))}{d(s, s')}, \quad (1.3)$$

whether or not these are finite. If, for instance,  $S$  is a real interval (with  $d(s, s') = |s - s'|$ ) and  $\psi$  is differentiable throughout  $S$ ,  $m(\psi)$  is the supremum of  $|\psi'(s)|$ . The hypothesis

H6  $m(\varphi_s) < \infty$  for all  $s \in E$

is a mere regularity condition. The next two assumptions, however, are genuinely restrictive:

H7  $\mu(f_e) \leq 1$  for all  $e \in E$ , and

H8 for any  $s \in S$  there is a positive integer  $k$  and there are  $k$  events  $e_1, \dots, e_k$  such that

$$\mu(f_{e_1 \dots e_k}) < 1 \text{ and } \phi_{e_1 \dots e_k}(s) > 0$$

where

$$\phi_{e_1 \dots e_k}(s) = P_s(E_j = e_j, 1 \leq j \leq k). \quad (1.4)$$

In H8 it is understood that the integers and events associated with different states may be different.

The inequality

$$d(g(s), g(s')) \leq \mu(g)d(s, s') \quad (1.5)$$

for mappings  $g$  of  $S$  into  $S$  suggests that such a function be called distance diminishing if  $\mu(g) \leq 1$  and strictly distance diminishing if  $\mu(g) < 1$ . Hypothesis H7 then says that all event operators are distance diminishing, while H8 says that, whatever the present state, some finite sequence of events with strictly distance diminishing cumulative effect can occur on subsequent trials. Both H7 and H8 (with  $k = 1$  for all states), are satisfied, for example, if all event operators are strictly distance diminishing.

It is now possible to introduce the following precise and convenient terminology.

Definition 1.1. A system  $((S, d), E, f, \varphi)$  of sets and functions is a distance diminishing model (or simply a model) if  $f(\cdot)$  maps  $E \times S$  into  $S$ ,  $\varphi(\cdot)$  maps  $E \times S$  into the non-negative real numbers,  $\sum_{e \in E} \varphi_e(s) = 1$ , and H2, H4, H5, H6, H7 and H8 are satisfied.

Definition 1.2. Stochastic processes  $\{s_n\}$  and  $\{E_n\}$  in the spaces  $S$  and  $E$ , respectively, are associated with the model if they satisfy H1 and H3.

## 2. SURVEY OF RESULTS

Most remarks on earlier work by other authors will be deferred until Section 4.

### a. Theorems Concerning States

The process  $\{S_n\}$  associated with any distance diminishing model is a Markov process with stationary transition probabilities given by

$$K(s, A) = \sum_{e: f_e(s) \in A} \varphi_e(s) = P_s(S_2 \in A) \quad (2.1)$$

for (Borel) subsets  $A$  of  $S$ . The  $n$  step transition probabilities for the process are given by

$$K^{(n)}(s, A) = \sum_{\substack{e_1 \dots e_n: \\ f_{e_1 \dots e_n}(s) \in A}} \varphi_{e_1 \dots e_n}(s) = P_s(S_{n+1} \in A), \quad (2.2)$$

for  $n \geq 1$ . It is convenient to let  $K^{(0)}(s, A)$  be 1 if  $s \in A$  and 0 otherwise. Functions like  $K$  and  $K^{(n)}$ , probability measures in their second variable for each value of their first, and measurable in their first variable for each value of their second, will be called stochastic kernels.

A basic problem is the asymptotic behavior of  $K^{(n)}(s, \cdot)$  as  $n \rightarrow \infty$ . Before considering this question, it is necessary to

specify what is meant by "convergence" of a sequence  $(\mu_n)$  of probability measures on  $S$  to a probability measure  $\mu$  on  $S$ . The appropriate notion is this:  $\mu_n$  converges to  $\mu$  if for any Borel subset  $A$  of  $S$

$$\mu(\overset{\circ}{A}) \leq \liminf_{n \rightarrow \infty} \mu_n(A) \text{ and } \limsup_{n \rightarrow \infty} \mu_n(A) \leq \mu(\bar{A})$$

where  $\overset{\circ}{A}$  is the interior and  $\bar{A}$  is the closure of  $A$ . If, for instance,  $S$  is a real interval, such convergence is equivalent to convergence of distribution functions at all points of continuity of the limit -- the usual notion of convergence for distribution functions. The extension of this notion to stochastic kernels that will be used below is as follows.

**Definition 2.1.** A sequence  $(K_n)$  of stochastic kernels converges uniformly to a stochastic kernel  $K_\infty$  if, for any Borel subset  $A$  of  $S$  and any  $\epsilon > 0$ , there is an integer  $N$  such that

$$K_\infty(s, \overset{\circ}{A}) - \epsilon \leq K_n(s, A) \leq K_\infty(s, \bar{A}) + \epsilon$$

for all  $n \geq N$  and  $s$  in  $S$ .

If a limiting stochastic kernel  $K_\infty(s, A)$  is independent of  $s$  for all  $A$ , it is sometimes natural to write  $K_\infty(A)$  instead of  $K_\infty(s, A)$ . Aside from this change of notation Definition 2.1 is unaffected.

A closely related problem is the asymptotic behavior of functions  $E[\varphi(S_n)]$  (moments, for instance) where  $\varphi$  is a real valued function on  $S$ . Two notions of convergence for sequences of real valued functions on  $S$  are important in what follows. For any such function  $\varphi$  define  $\|\varphi\|$  and  $\|\varphi\|_1$  by

$$|\gamma| = \sup_{s \in S} |\gamma(s)| \text{ and} \quad (2.3)$$

$$\|\gamma\| = |\gamma| + m(\gamma). \quad (2.4)$$

The class of continuous real valued functions on  $S$  is denoted  $C(S)$  (note that  $|\gamma| < \infty$  if  $\gamma \in C(S)$ ), and the subclass on which  $m(\gamma) < \infty$  (and thus  $\|\gamma\| < \infty$ ) is denoted  $CL$ . A sequence  $\{\gamma_n\}$  of functions in  $C(S)$  converges uniformly to  $\gamma \in C(S)$  if  $|\gamma_n - \gamma| \rightarrow 0$  as  $n \rightarrow \infty$ . A stronger notion of convergence, applicable to functions in  $CL$ , is  $\|\gamma_n - \gamma\| \rightarrow 0$  as  $n \rightarrow \infty$ . If  $S$  is a real interval then the collection  $D$  of functions with a bounded derivative is a closed subset of  $CL$  in the sense that, if  $\gamma_n \in D$ ,  $\gamma \in CL$ , and  $\lim_{n \rightarrow \infty} \|\gamma_n - \gamma\| = 0$ , then  $\gamma \in D$ . Since  $\|\psi\| = |\psi| + |\psi'|$  for any  $\psi \in D$ , it follows that  $|\gamma_n - \gamma| \rightarrow 0$  and  $|\gamma'_n - \gamma'| \rightarrow 0$  as  $n \rightarrow \infty$ . If  $f_0$  and  $\psi_0 \in D$  for all  $s \in S$  and if  $\psi \in D$ , then  $E_n[\psi(s_n)] \in D$  for all  $n \geq 1$ . Thus these observations are applicable to  $\gamma_n(\cdot) = E_n[\psi(s_n)]$  and  $\gamma_n(\cdot) = (1/n) \sum_{j=1}^n E_n[\psi(s_j)]$ .

Theorem 2.1 gives some information about the asymptotic behavior of  $\{s_n\}$  for distance diminishing models with no further assumptions.

Theorem 2.1. For any distance diminishing model, the stochastic kernel  $(1/n) \sum_{j=0}^{n-1} K^{(j)}$  converges uniformly as  $n \rightarrow \infty$  to a stochastic kernel  $K^*$ . For any Borel subset  $A$  of  $S$ ,  $K^*(\cdot, A) \in CL$ . There is a <sup>constant</sup>  $L_A < \infty$  such that

$$\left\| \frac{1}{n} \sum_{j=1}^n E_n[\psi(s_j)] - E_n[\psi(s_n)] \right\| \leq c \frac{\|\psi\|}{n} \quad (2.5)$$

for all  $n \geq 1$  and  $\psi \in CL$ , where

$$E_s[\psi(S_\infty)] = \int_S \psi(s') K^\infty(s, ds'). \quad (2.6)$$

The notation  $E_s[\psi(S_\infty)]$  for the expectation of  $\psi$  with respect to the asymptotic distribution  $K^\infty(s, \cdot)$  is not meant to suggest that there is a random variable  $S_\infty$  to which  $S_n$  converges with probability 1. Though such convergence occurs, for example, under the hypotheses of Theorem 2.3, it does not occur in general.

Two situations will now be discussed in which the conclusions of Theorem 2.1 can be substantially strengthened. The first is characterized by the loss, asymptotically, of all information about the initial state; the second, by the convergence of  $S_n$  to absorbing states with probability 1. Both occur frequently in mathematical learning theory. To describe hypotheses that lead to these situations, it is convenient to have a notation for the set of values that  $S_{n+1}$  takes on with positive probability when  $S_1 = s$ . This set is denoted  $T_n(s)$ :

$$T_n(s) = \{s' : K^{(n)}(s, \{s'\}) > 0\}. \quad (2.7)$$

An absorbing state is, of course, one that, once entered, cannot be left; that is,  $K(s, \{s\}) = 1$  or  $T_1(s) = \{s\}$ . Another convenient notation is  $d(A, B)$  for the (minimum) distance between two subsets  $A$  and  $B$  of  $S$ :

$$d(A, B) = \inf_{s \in A, s' \in B} d(s, s') \quad (2.8)$$

If  $B$  is the unit set  $\{b\}$ , then  $d(A, B)$  is written  $d(A, b)$ .

Theorem 2.2 shows that, to obtain asymptotic independence of the initial state, it suffices to assume that

$$H9 \quad \lim_{n \rightarrow \infty} d(T_n(s), T_n(s')) = 0 \text{ for all } s, s' \in S.$$

Theorem 2.3 shows that, to obtain convergence to absorbing states, it suffices to assume that:

H10 There are a finite number of absorbing states  $a_1, \dots, a_N$ , such that, for any  $s \in S$ , there is some  $a_j(s)$  for which

$$\lim_{n \rightarrow \infty} d(T_n(s), a_j(s)) = 0.$$

It is easy to see that H9 and H10 are inconsistent except when there is exactly one absorbing state, in which case they are equivalent.

Theorem 2.2. If a distance diminishing model satisfies H9, then the asymptotic distribution  $K^\infty(s, \cdot) = K^\infty(\cdot)$  does not depend on the initial state  $s$ , and  $K^{(n)}$  converges uniformly to  $K^\infty$ . There are constants  $\alpha < 1$  and  $C < \infty$  such that

$$\|E[\psi(S_n)] - E[\psi(S_\infty)]\| \leq C\alpha^n \|\psi\| \quad (2.9)$$

for  $n \geq 1$  and  $\psi \in CL$ , where  $E[\psi(S_\infty)] = \int_S \psi(s) K^\infty(ds)$ .

**Theorem 2.3.** If a distance diminishing model satisfies H10,  
then the stochastic process  $\{S_n\}$  converges with probability 1 to  
a random absorbing state  $S_\infty$ . For any  $1 \leq i \leq N$ , the function  
 $\gamma_i(s) = P(S_\infty = a_i | S_0 = s)$  belongs to CL. If  $b_1, \dots, b_N$  are real numbers,  
the function  $\gamma(\cdot) = \sum_{i=1}^N b_i \gamma_i(s)$  is the  
only continuous  
solution of the equation  $E_s[\gamma(S_2)] = \gamma(s)$  that has the boundary  
values  $\gamma(a_j) = b_j$ . The stochastic kernels  $K^{(n)}$  converge uni-  
formly to  $K^\infty$ , and  $K^\infty(s, \cdot)$  assigns weight  $\gamma_i(s)$  to  $a_i$ , so that  
 $E_s[\psi(S_\infty)] = \sum_{i=1}^N \gamma_i(s) \psi(a_i)$ . There are  $\alpha < 1$  and  $C < \infty$  such that

$$\|E_s[\psi(S_n)] - E_s[\psi(S_\infty)]\| \leq C\alpha^n \|\psi\| \quad (2.10)$$

for all  $n \geq 1$  and  $\psi \in CL$ .

These theorems suggest the following terminology:

**Definition 2.2.** A distance diminishing model is ergodic if  
it satisfies H9, and absorbing if it satisfies H10.

Note that, whereas in Theorem 2.1 only the convergence of Cesaro averages is asserted, in Theorems 2.2 and 2.3 the sequences  $\{K^{(n)}\}$  and  $\{E[\psi(S_n)]\}$  themselves converge. It is also worth pointing out that, although it is of little importance that (2.9) and (2.10) imply  $\|E_s[\psi(S_n)] - E_s[\psi(S_\infty)]\| \rightarrow 0$  instead of simply  $|E_s[\psi(S_n)] - E_s[\psi(S_\infty)]| \rightarrow 0$ , it is of considerable importance that these formulas give a geometric rate of convergence, independent of  $\psi$  as long as  $\|\psi\|$  is less than some fixed constant.

Proofs of Theorems 2.1 - 2.3 are given in Section 5. The main tool used is the uniform ergodic theorem of Ionescu Tulcea and Marinescu (1950). The results given above do not exhaust

the implications of this theorem, even for distance diminishing models, as will be seen in Section 5.

b. Theorems Concerning Events

consider some characteristic  $C^l$  that pertains to  $l$  consecutive events,  $l \geq 1$ ; e.g., "response R occurs on trial  $n$ " ( $l = 1$ ), "the responses on trials  $n$  and  $n+1$  differ" ( $l = 2$ ), or "outcome O occurs on trial  $n$  and response R on trial  $n+1$ " ( $l = 2$ ). It is often of interest to know the asymptotic behavior of the probability that  $(E_n, \dots, E_{n+l-1})$  has the property  $C^l$ . Let  $E^l$  be the set of  $n$  tuples of events, and let  $A^l$  be the subset of  $E^l$  that corresponds to  $C^l$ ; that is,  $A^l = \{(e_1, \dots, e_l) : (e_1, \dots, e_l) \text{ has the property } C^l\}$ . Then it is the asymptotic behavior of

$$P_S^{(n)}(A^l) = P_S((E_n, \dots, E_{n+l-1}) \in A^l) \quad (2.11)$$

that is in question. Theorem 2.4, which applies to both ergodic and absorbing models, gives much information.

Theorem 2.4. For any ergodic or absorbing model there is an  $L < \infty$  such that, for any  $l \geq 1$  and  $A^l \subset E^l$ ,

$$\|P_S^{(n)}(A^l) - P_S^\infty(A^l)\| \leq L\alpha^n \quad (2.12)$$

for all  $n \geq 1$ , where

$$P_S^\infty(A^l) = \int_S P_{S'}^{(1)}(A^l) K^\infty(s, ds'), \quad (2.13)$$

and  $\alpha$  is as in (2.9) and (2.10).

In the ergodic case the subscript  $s$  on  $P_s^{\infty}(A^L)$  can, of course, be dropped.

The following corollary for absorbing models is very useful.

Corollary 2.5. If an absorbing model and an  $A^L \subset E^L$  have the property that  $P_{a_i}^{(1)}(A^L) = 0$  for  $i = 1, \dots, N$ , then the total number  $X$  of positive integers  $n$  for which  $(E_n, \dots, E_{n+L-1}) \in A^L$  is finite with probability 1, and

$$\|E_s[X]\| \leq La/(1 - \alpha). \quad (2.14)$$

The function  $\chi(s) = E_s[X]$  is the unique continuous solution of the equation

$$\chi(s) = P_s^{(1)}(A^L) + E_s[\chi(S_2)]$$

for which  $\chi(a_i) = 0$ ,  $i = 1, \dots, N$ .

The next theorem concerns ergodic models, and requires some additional notation for its statement. Let  $h$  be a real valued function on  $E$ . Then the asymptotic expectations of  $h(E_n)$  and  $h(E_n)h(E_{n+j})$  are denoted  $E[h(E_n)]$  and  $E[h(E_n)h(E_{n+j})]$ , respectively.

Thus

$$E[h(E_n)] = \sum_{e \in E} h(e)P^{\infty}(\{e\}), \text{ and} \quad (2.15)$$

$$E[h(E_n)h(E_{n+j})] = \sum_{e_1, \dots, e_{j+1}} h(e_1)h(e_{j+1})P^{\infty}(\{(e_1, \dots, e_{j+1})\}). \quad (2.16)$$

In typical applications  $h$  will be the indicator function of some  $A \subset E$ , so that  $\sum_{j=m}^{m+n-1} h(E_j)$  is the number of occurrences of events in  $A$  during the block of  $n$  trials beginning on trial  $m$ . In this case,

$$E[h(E_\infty)] = P^\infty(A),$$

$$E[h(E_\infty)h(E_{\infty+1})] = P^\infty(A \times A), \text{ and}$$

$$E[h(E_\infty)h(E_{\infty+j})] = P^\infty(A \times E^{j-1} \times A) \text{ for } j \geq 2.$$

Theorem 2.6.(i) For any ergodic model, and any real valued function  $h$  on  $E$ , the series

$$E[h^2(E_\infty)] - E^2[h(E_\infty)] + 2 \sum_{j=1}^{\infty} (E[h(E_\infty)h(E_{\infty+j})] - E^2[h(E_\infty)]) \quad (2.17)$$

converges to a non-negative constant  $\sigma_h^2$ .

(ii) For some  $C_h < \infty$  and all  $m, n \geq 1$

$$\left| \frac{1}{n} E \left[ \left( \sum_{j=m}^{m+n-1} h(E_j) - nE[h(E_\infty)] \right)^2 \right] - \sigma_h^2 \right| \leq C_h n^{-1/2}. \quad (2.18)$$

Consequently, the law of large numbers

$$\lim_{n \rightarrow \infty} E \left[ \left( \frac{1}{n} \sum_{j=m}^{m+n-1} h(E_j) - E[h(E_\infty)] \right)^2 \right] = 0 \quad (2.19)$$

holds uniformly in  $s$ .

(iii) If  $\sigma_h^2 > 0$ , the central limit theorem

$$\lim_{n \rightarrow \infty} P_s \left( \frac{\sum_{j=1}^{n+n-1} h(\mathbb{E}_j) - nE[h(\mathbb{E}_s)]}{\sqrt{n} \sigma_h} < x \right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp(-t^2/2) dt \quad (2.20)$$

is valid for all  $s \in S$ .

A distance diminishing model can be regarded as an example of what Iosifescu (1963) calls a homogeneous random system with complete connections. Theorem 2.6 is a consequence of Theorem 2.4 and a theorem of Iosifescu on such systems. Results in this subsection will be proved in Section 6.

## 3. EXAMPLES

The examples to be discussed have been selected so as to illustrate a variety of ramifications of the theory developed in Sections 1 and 2.

a. Linear Models with Experimenter-Subject Controlled Events

Suppose that the subject in a learning experiment has response alternatives  $A_1$  and  $A_2$  on each trial, and that, following response  $A_i$ , one of two observable outcomes  $O_{i1}$  and  $O_{i2}$  occurs. It is assumed that  $O_{1j}$  and  $O_{2j}$  positively reinforce  $A_j$ , in the weak sense that they do not decrease the probability of  $A_j$ . The outcome probabilities are supposed to depend at most on the most recent response. Let  $A_{i,n}$  and  $O_{ij,n}$  denote, respectively, the occurrence of  $A_i$  and  $O_{ij}$  on trial  $n$ , and denote the probability  $P(O_{ij,n} | A_{i,n})$  by  $\pi_{ij}$ .

Linear models with experimenter-subject controlled events (Bush and Mosteller (1955)) for this situation can be described within the framework of Section 1 by identifying  $p_n$ , the (conditional) probability of  $A_{1,n}$ , with the state  $S_n$ , by identifying the response-outcome pair that occurs on trial  $n$  with the event  $E_n$ , and by making the following stipulations:

$$S = [0, 1], d(p, p') = |p - p'|, \quad (3.1)$$

$$(i, j) = (A_i, O_{ij}) \text{ and } E = \{(i, j) : 1 \leq i, j \leq 2\}, \quad (3.2)$$

$$f_{ij}(p) = (1 - \theta_{ij})p + \theta_{ij}\delta_{j1}, \quad (3.3)$$

$$\varphi_{ij}(p) = (p\delta_{i1} + (1 - p)\delta_{i2})\tau_{ij}, \quad (3.4)$$

$$\tau_{i1} + \tau_{i2} = 1, \text{ and } 0 \leq \theta_{ij}, \tau_{ij} \leq 1 \text{ for } 1 \leq i, j \leq 2. \quad (3.5)$$

In (3.3) and (3.4),  $\delta_{ij}$  is the Kronecker  $\delta$ . For convenience, any system  $((S, d), E, f, \varphi)$  of sets and functions satisfying (3.1) - (3.5) will be referred to as a four-operator model. In this terminology, (3.1) - (3.5) define a six-parameter family of four-operator models, one for each choice of  $\theta_{11}, \theta_{12}, \theta_{21}, \theta_{22}, \tau_{11}$ , and  $\tau_{22}$  consistent with (3.5). Since  $m(\varphi_{ij}) = \tau_{ij}$  and  $\mu(f_{ij}) = (1 - \theta_{ij}) \leq 1$ , it is clear that any four-operator model satisfies all of the conditions of Definition 1.1 except perhaps H3.

The asymptotic behavior of the process  $(p_n)$  associated with a four-operator model depends critically on the number of absorbing states. Lemma 3.1 catalogues the absorbing states for a four-operator model.

**Lemma 3.1.** The state 1 is absorbing if and only if  $\tau_{12} = 0$  or  $\theta_{12} = 0$ . The state 2 is absorbing if and only if  $\tau_{21} = 0$  or  $\theta_{21} = 0$ . A state  $p \in (0, 1)$  is absorbing if and only if for each  $(i, j) \in E$ ,  $\theta_{ij} = 1$  or  $\tau_{ij} = 0$ . In this case all states are absorbing, and the model is said to be trivial.

**Proof.** A state  $p \in (0, 1)$  is absorbing if and only if, for any  $(i, j) \in E$ , either  $f_{ij}(p) = p$  (in which case  $\theta_{ij} = 0$  and  $f_{ij}(x) = x$ ) or  $\varphi_{ij}(p) = 0$  (in which case  $r_{ij} = 0$  and  $\varphi_{ij}(x) = 0$ ).

The state 1 is absorbing if and only if  $1 - \theta_{12} = f_{12}(1) = 1$  or  $r_{12} = \varphi_{12}(1) = 0$ . The assertion concerning the state 0 is proved similarly. Q.E.D.

The next lemma tells which four-operator models satisfy NS.

**Lemma 3.2.** A four-operator model is distance diminishing if and only if, for each  $i \in \{1, 2\}$ , there is some  $j_1 \in \{1, 2\}$  such that  $\theta_{ij_1} > 0$  and  $r_{ij_1} > 0$ .

**Proof.** Suppose that the condition given by the lemma is met. If  $p > 0$  then  $\varphi_{1j_1}(p) = pr_{1j_1} > 0$  and  $\mu(f_{1j_1}) = 1 - \theta_{1j_1} < 1$ . Similarly, if  $p < 1$  then  $\varphi_{2j_2}(p) > 0$  and  $\mu(f_{2j_2}) < 1$ . Thus NS is satisfied with  $k = 1$  for all states.

Suppose that the condition fails. Then for some  $i \in \{1, 2\}$  and all  $j \in \{1, 2\}$ ,  $\theta_{ij} = 0$ , or  $r_{ij} = 0$ . Since the cases  $i = 1$  and  $i = 2$  can be treated similarly, only  $i = 1$  will be considered.

It follows from Lemma 3.1 on taking  $j = 2$  that 1 is an absorbing state. Thus  $\varphi_{1n_1}, \dots, \varphi_{1n_k}(1) > 0$  implies  $n_j = 1$  and  $r_{1n_j} > 0$ ,  $1 \leq j \leq k$ . But then  $\theta_{n_j} = 0$  for  $1 \leq j \leq k$  and  $\mu(f_{n_1, \dots, n_k}) = 1$ . Q.E.D.  
C. NS is not satisfied.

Clearly a distance diminishing four-operator model is non-trivial.

With one inconsequential exception, distance diminishing four operator models are either ergodic or absorbing. Theorems 3.1 and 3.2 show slightly more.

Theorem 3.1. If neither 0 nor 1 is absorbing for a four-operator model, then  $\theta_{ij} > 0$  and  $r_{ij} > 0$  for  $i \neq j$ , and the model is distance diminishing. Either (i)  $\theta_{ij} = 1$  and  $r_{ij} = 1$  if  $i \neq j$ ; or (ii) the model is ergodic.

Theorem 3.2. If a distance diminishing four-operator model has an absorbing state, then it is an absorbing model.

The behavior of the process  $\{p_n\}$  when  $\theta_{ij} = 1$  and  $r_{ij} = 1$  for  $i \neq j$  is completely transparent. Starting at  $p$  the process moves on its first step to 1 with probability  $1 - p$  and to 0 with probability  $p$ , and thereafter alternates between these two extreme

states. This cyclic model is of no psychological interest and will be discussed no further.

Proof of Theorem 3.1. By Lemma 3.1 if neither 0 nor 1 is absorbing then  $\theta_{ij} > 0$  and  $\pi_{ij} > 0$  for  $i \neq j$ , and the model is distance diminishing by Lemma 3.2.

Suppose  $\pi_{21} < 1$ . Then by considering first the case  $p = 0$ , then  $p > 0$  and  $\theta_{12} = 1$ , and finally  $p > 0$  and  $\theta_{12} < 1$ , it is seen that  $(1 - \theta_{12})^n p \in T_n(p)$  for all  $n \geq 1$ . Thus  $d(T_n(p), T_n(q)) \leq (1 - \theta_{12})^{n-1} |p - q| \rightarrow 0$  as  $n \rightarrow \infty$ , and the model is ergodic according to Definition 2.1. By symmetry the same conclusion obtains if  $\pi_{12} < 1$ . Suppose that  $\theta_{12} < 1$ . Then  $(1 - \theta_{12})^n p \in T_n(p)$  for all  $p > 0$  and  $n \geq 1$ , and  $(1 - \theta_{12})^{n-1} \theta_{21} \in T_n(0)$  for all  $n \geq 1$ . Since both sequences tend to 0, ergodicity follows. The same conclusion follows by symmetry when  $\theta_{21} < 1$ . Thus if (1) does not hold the model is ergodic. Q.E.D.

Proof of Theorem 3.2. The condition given by Lemma 3.2 for a four-operator model to be distance diminishing allows four possibilities. These are distinguished by the values of  $j_i$ ,  $i = 1, 2$ : A:  $j_1 = 1, j_2 = 1$ ; B:  $j_1 = 2, j_2 = 2$ ; C:  $j_1 = 1, j_2 = 2$ ; and D:  $j_1 = 2, j_2 = 1$ . Lemma 3.1 shows that D is inconsistent with the existence of absorbing states. Thus it remains to show that a model is absorbing under A, B, or C if there are absorbing states.

Under A,  $1 - (1 - \theta_{21})^n (1 - p) \in T_n(p)$  for all  $n \geq 1$  and  $0 \leq p \leq 1$ , so  $d(T_n(p), 1) \leq (1 - \theta_{21})^n \rightarrow 0$  as  $n \rightarrow \infty$ . This implies that 0 is not an absorbing state. By assumption, however, there

is at least one absorbing state, so 1 is absorbing. But then  $\lim_{n \rightarrow \infty} d(T_n(p), 1) = 0$  for all  $0 \leq p \leq 1$  implies that the model is absorbing. By symmetry the model is also absorbing under B.

If 0 is not absorbing  $\tau_{21} > 0$  and  $\theta_{21} > 0$  by Lemma 3.1. Thus, if C holds, A does also, and the model is absorbing. If C holds, and 1 is not absorbing, the same conclusion follows by symmetry. Condition C implies that  $(1 - \theta_{22})^n p \in T_n(p)$  for  $p < 1$ ,  $1 - (1 - \theta_{11})^n (1 - p) \in T_n(p)$  for  $p > 0$ , and  $\theta_{11}, \theta_{22} > 0$ . Thus if both 0 and 1 are absorbing, H0 is satisfied with  $j(1) = 1$ ,  $j(0) = 0$  and  $j(p) = 1$  or 0 for  $0 < p < 1$ . Q.E.D.

As a consequence of Theorems 3.1 and 3.2 all of the theorems of Section 2 for ergodic models are valid for non-cyclic four-operator models without absorbing states, and all theorems of Section 2 for absorbing models are valid for distance diminishing four-operator models with absorbing states. A few illustrative specializations of the theorem of Section 2 to the case at hand will now be given. The first concerns convergence of the moments  $E_p[p_n^v]$  of the process  $\{p_n\}$ .

Theorem 3.3. For any non-cyclic distance diminishing four-operator model there are constants  $C < \infty$  and  $\alpha < 1$  such that

$$\|E_p[p_n^v] - E_p[p_\infty^v]\| \leq C(v+1)\alpha^n \quad (3.6)$$

for all real  $v \geq 1$  and positive integers  $n$ . The function  $E_p[p_\infty^v]$  has a bounded derivative.

This is obtained from (2.9) and (2.10) by noting that the function  $\psi(p) = p^v$  belongs to  $D$  with  $|\psi| = 1$  and  $m(\psi) = |\psi^2| = v$ , so that  $\|\psi\| = v + 1$ .

If 0 is the only absorbing state of a distance diminishing four-operator model, Theorem 2.3 implies that  $\lim_{n \rightarrow \infty} p_n = 0$  with probability 1, whatever the value of  $p_1$ . It is conceivable, however, that the convergence is sufficiently slow that the total number  $X$  of  $A_1$  responses is infinite with positive probability. Furthermore, even if  $X$  is finite with probability 1, it might have an infinite mean. Similarly, even though  $p_n$  converges to 0 or 1 in the case of two absorbing states, a priori considerations do not rule out the possibility that the total number  $Y$  of alternations between responses is infinite, or barring that, that its mean is infinite. Theorem 3.4 excludes these possibilities.

Theorem 3.4. If 0 is the only absorbing state of a distance diminishing four-operator model, then  $X$ , the total number of  $A_1$  responses, is finite with probability 1, and  $\|E[X]\| < \infty$ . If both 0 and 1 are absorbing states of a distance diminishing four-operator model, then  $Y$ , the total number of alternations between responses, is finite with probability 1 and  $\|E[Y]\| < \infty$ .

Naturally the first assertion is still true if 1 replaces 0 as the only absorbing state and  $X$  is the total number of  $A_2$  responses.

Proof. Let  $B$  and  $D$  be the subsets of  $E$  and  $E^2$ , respectively, defined by

$$B = \{(1, 1), (1, 2)\} \text{ and} \quad (3.7)$$

$$D = \{((i, j), (k, l)): i \neq k\}.$$

Then  $E_n \in B$  if and only if  $A_1$  occurs on trial  $n$ , and  $(E_n, E_{n+1}) \in D$  if and only if there is a response alternation between trials  $n$  and  $n+1$ . Since  $P_0^{(1)}(B) = 0$ , and, if both 0 and 1 are absorbing,  $P_0^{(1)}(D) = P_1^{(1)}(D) = 0$ , the conclusions of the theorem follow directly from Corollary 2.5. Q.E.D.

If  $h$  is the indicator function of the subset  $B$  of  $E$  given by (3.7), then  $A_n = h(E_n)$  is the indicator random variable of  $A_{1,n}$  and  $\sum_{j=m}^{m+n-1} A_j$  is the frequency of  $A_1$  in the block of  $n$  trials beginning on trial  $m$ . Theorems 3.1 and 2.6 yield a law of large numbers and, perhaps, a central limit theorem for this quantity for any non-cyclic four-operator model with no absorbing states. The full power of this result comes into play when the quantities  $E[A_\infty] = \lim_{n \rightarrow \infty} P(A_{1,n})$  and  $\sigma_h^2$  can be computed explicitly in terms of the parameters of the model. This is the case, for instance, when all  $\theta_{ij}$  are equal.

Theorem 3.5. A four-operator model with  $\theta_{ij} = \theta > 0$  for  $1 \leq i, j \leq 2$ , and  $\pi_{ij} > 0$  for  $i \neq j$ , but not  $\theta = \pi_{12} = \pi_{21} = 1$ , is ergodic. The law of large numbers

$$\lim_{n \rightarrow \infty} E_p [((1/n)S_{m,n} - l)^2] = 0 \quad (3.8)$$

and central limit theorem

$$\lim_{n \rightarrow \infty} P_p \left( \frac{S_{m,n} - nl}{\sqrt{n} \sigma} < x \right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp(-t^2/2) dt, \quad (3.9)$$

hold, where

$$l = \pi_{21}/(\pi_{21} + \pi_{12}), \quad (3.10)$$

$$\sigma^2 = \frac{l(1-l)}{\pi_{21} + \pi_{12}} \left[ \pi_{11} + \pi_{22} + \frac{2(1-\theta)}{(2-\theta) + 2(1-\pi_{11}-\pi_{22})(1-\theta)} \right], \quad (3.11)$$

and  $S_{m,n}$  is the total number of  $A_1$  responses in the  $n$  trial block beginning on trial  $m$ .

Outline of proof. Ergodicity follows from Theorem 3.1, so Theorem 2.6 is applicable. Straightforward computation yields

$$E[p_\infty] = E[A_\infty] = l, \quad (3.12)$$

$$E[p_\infty^2] = l^2 + \theta l(1-l)/[(2-\theta) + 2(1-\pi_{11}-\pi_{22})(1-\theta)], \quad (3.13)$$

$$E[A_\infty A_{\infty+1}] = (1-\theta)E[p_\infty^2] + \pi_{11}\theta E[p_\infty], \text{ and} \quad (3.14)$$

$$E[A_\infty A_{\infty+k}] - E^2[A_\infty] = (1-\theta(\pi_{12} + \pi_{21}))^{k-1} (E[A_{\infty+1} A_\infty] - E^2[A_\infty]) \quad (3.15)$$

for  $k \geq 1$ . These formulas permit computation of  $\sigma^2 = \sigma_n^2$ , the series in (2.17). The result, recorded in (3.11), is positive, since  $0 < \ell < 1$ , and either  $\pi_{11} + \pi_{22} > 0$  or  $(1 - \theta) > 0$ . Q.E.D.

The equality

$$M(p) - p = (\pi_{12} + \pi_{21})(\ell - p),$$

where

$$M(p) = P(O_{11,n} \text{ or } O_{21,n} | p_n = p),$$

shows that the asymptote  $\ell$  of  $A_1$  response probability for the linear model with equal  $\theta$ 's is associated with the asymptotic equality of the probability of  $A_1$  and the probability of reinforcement of  $A_1$ :

$$\lim_{n \rightarrow \infty} P(O_{11,n} \text{ or } O_{21,n}) = \lim_{n \rightarrow \infty} P(A_{1,n}).$$

Such probability matching is a well known prediction of the linear model with equal  $\theta$ 's. Theorem 3.5 contains a much stronger prediction. The law of large numbers (3.8) asserts that the proportion  $(1/n) S_{m,n}$  of  $A_1$  responses for a single subject in a long block of trials is close to  $\ell$  with high probability. The terms "close" and "high" are further quantified by the central limit theorem (3.9). To illustrate, if reinforcement is noncontingent with  $\pi_{11} = \pi_{21} = .75$  and  $\theta$  is small (that is, learning is slow), then  $\sigma^2 \approx 2\pi(1 - \pi) = .375$  so that, in a block of 400 trials commencing on trial 100, the

probability is approximately .01 that  $(1/400)s_{100,400}$  will depart from .75 by as much as  $(2.58)(.612)/20 = .079$ .

There is one modification of the four-operator model examples of which have occurred sufficiently frequently in the literature (see Estes and Suppes (1959), Norman (1964), and Yellott (1965)) to warrant comment here. If, following any of the outcomes  $0_{ij}$ , conditioning is assumed to be effective (i.e.,  $p_{n+1} = f_{ij}(p_n)$ ) with probability  $c_{ij}$  and, otherwise, ineffective (i.e.,  $p_{n+1} = p_n$ ) a five-operator model is obtained. It is easy to amend (3.2) - (3.5) to obtain a formal description within the framework of Section 1. Such an addition of an identity operator does not affect the validity of any of the results preceding Theorem 3.5 (or their proofs) provided that  $\pi_{ij}$  is everywhere replaced by  $\pi_{ij}c_{ij}$ . The first sentence of the amended Theorem 3.5 should read: A five-operator model with  $\theta_{ij} = \theta > 0$  and  $c_{ij} = c > 0$  for  $1 \leq i, j \leq 2$ , and  $\pi_{ij} > 0$  for  $i \neq j$ , but not  $\theta = c = \pi_{12} = \pi_{21} = 1$ , is ergodic. Also (3.1) should be replaced by

$$\sigma^2 = \frac{\ell(1-\ell)}{\pi_{21} + \pi_{12}} \left[ \pi_{11} + \pi_{22} + \frac{2(1-c\theta)}{c((2-\theta) + 2(1-\pi_{11}-\pi_{22})(1-\theta))} \right], \quad (3.16)$$

and  $\theta$  should be replaced by  $c\theta$  in (3.14) and (3.15). An interesting implication of (3.16) is that  $\lim_{\theta \rightarrow 0} \sigma^2 < \infty$ , whereas, if  $c\theta < 1$ ,  $\lim_{c \rightarrow 0} \sigma^2 = \infty$ . Thus the variance of the total number of  $A_1$  responses in a long block of trials may be useful in deciding whether a given instance of "slow learning" is due to small  $\theta$  or small  $c$ .

b. Lovejoy's Model I

Lovejoy's (1966) Model I is a simple model for simultaneous discrimination learning. Let the relevant stimulus dimension be brightness, and let white (W) be positive and black (B) be negative. On each trial the subject is supposed either to attend to brightness (A) or not ( $\tilde{A}$ ), which events have probabilities  $P_n(A)$  and  $1 - P_n(A)$ . Given A the probability of the response appropriate to white is  $P_n(W|A)$ , while given  $\tilde{A}$  the probability of this response is  $1/2$ . The subject's state on trial n is then described by the vector  $(P_n(A), P_n(W|A))$ , and the state space is

$$S = \{(p, p') : 0 \leq p, p' \leq 1\}.$$

This is a compact metric space with respect to the ordinary Euclidean metric d.

The events are the elements of

$$E = \{(A, W), (A, B), (\tilde{A}, W), (\tilde{A}, B)\},$$

the corresponding transformations are

$$f_{AW}(p, p') = (\alpha_1 p + (1 - \alpha_1), \alpha_3 p' + (1 - \alpha_3)),$$

$$f_{AB}(p, p') = (\alpha_2 p, \alpha_4 p' + (1 - \alpha_4)),$$

$$f_{\lambda W}(p, p') = (\alpha_1 p, p'), \text{ and}$$

$$f_{\lambda B}(p, p') = (\alpha_2 p + (1 - \alpha_2), p'),$$

where  $0 < \alpha_1, \alpha_2, \alpha_3, \alpha_4 < 1$ , and their probabilities are

$$\varphi_{\lambda W}(p, p') = pp',$$

$$\varphi_{\lambda B}(p, p') = p(1 - p'),$$

$$\varphi_{\lambda W}(p, p') = (1 - p)/2, \text{ and}$$

$$\varphi_{\lambda B}(p, p') = (1 - p)/2.$$

Any system  $((S, d), E, f, \varphi)$  of sets and functions satisfying the above stipulations will be called a discrimination model of type I below.

**Theorem 3.6.** Any discrimination model of type I is distance diminishing and absorbing with single absorbing state (1, 1).

**Proof.** Axiom H6 is satisfied because of the continuous differentiability of the  $\varphi$ 's, and H7 follows from

$$\mu(f_{\lambda W}) = \max(\alpha_1, \alpha_3) < 1,$$

$$\mu(f_{\lambda B}) = \max(\alpha_2, \alpha_4) < 1, \text{ and}$$

$$\mu(f_{\lambda W}) = \mu(f_{\lambda B}) = 1.$$

Thus it remains only to verify H8 and H10.

Note that, as a consequence of (1.5), for any mappings  $f$  and  $g$  of  $S$  into  $S$  such that  $\mu(f) < \infty$  and  $\mu(g) < \infty$ , the inequality

$$\mu(f \circ g) \leq \mu(f)\mu(g)$$

obtains, where  $f \circ g(s) = f(g(s))$ . This implies that  $\mu(f_{e_1 \dots e_k}) < 1$  if  $e_k = (A, W)$ . Now  $\varphi_{AW}(p, p') > 0$  throughout

$$S' = \{(p, p') : p > 0, p' > 0\},$$

so to complete the verification of H8 it suffices to show that if  $(p, p') \in S'$  there is a  $k \geq 2$  and there are events  $e_1, \dots, e_{k-1}$  such that  $f_{e_1 \dots e_{k-1}}(p, p') \in S'$  and  $\varphi_{e_1 \dots e_{k-1}}(p, p') > 0$ . If  $0 < p' \leq 1$ ,  $\varphi_{AB}(0, p') > 0$  and  $f_{AB}(0, p') \in S'$ , while if  $0 < p \leq 1$ ,  $\varphi_{AB}(p, 0) > 0$  and  $f_{AB}(p, 0) \in S'$ . Finally  $f_{AB}(0, 0)$  has positive first and null second coordinate, so  $f_{AB,AB}(0, 0) \in S'$  and  $\varphi_{AB}(f_{AB}(0, 0)) > 0$ . Since  $\varphi_{AB}(0, 0) > 0$  the latter inequality implies  $\varphi_{AB,AB}(0, 0) > 0$ .

The above argument shows that for any  $(p, p') \in S$  there is a point  $s^{PP'} \in T_{k-1}(p, p') \cap S'$ . Since  $f_{AW}$  maps  $S'$  into  $S'$  it follows that  $f_{AW}^{(n)}(s^{PP'}) \in T_{n+k-1}(p, p')$  for  $n \geq 0$ , where  $f_{AW}^{(n)}$  is the  $n^{\text{th}}$  iterate of  $f_{AW}$ , i.e.,

Norman

33.

$f_{AW}^{(0)}(s) = s$  and  $f_{AW}^{(j+1)} = f_{AW} \circ f_{AW}^{(j)}$ ,  $j \geq 0$ . But for any  $(q, q') \in S$  and  $n \geq 0$ ,

$$f_{AW}^{(n)}(q, q') = (1 - (1 - q)\alpha_1^n, 1 - (1 - q')\alpha_3^n),$$

so

$$d(T_{n+k-1}(p, p'), (1, 1)) \leq d(f_{AW}^{(n)}(s^{pp'}), (1, 1))$$

$$\leq (\alpha_1^{2n} + \alpha_3^{2n})^{1/2} \rightarrow 0 \text{ as } n \rightarrow \infty.$$

Since  $(1, 1)$  is obviously an absorbing state, the verification of H10 is complete. Q.E.D.

Here is a sample of what can be concluded about Lovejoy's

Model I on the basis of Theorem 3.6, Theorem 2.3, and Corollary 2.5.

discrimination

**Theorem 3.7.** For any model of type I,  $\lim_{n \rightarrow \infty} P_n(A) = 1$  and  $\lim_{n \rightarrow \infty} P_n(W|A) = 1$  with probability 1. There are constants  $C < \infty$  and  $\alpha < 1$  such that

$$\|E.[P_n^v(A)P_n^w(W|A)] - 1\| \leq C((v^2 + w^2)^{1/2} + 1)\alpha^n, \quad (3.17)$$

for all real  $v, w \geq 1$  and positive integers  $n$ . The total number  $Z$  of  $B$  responses is finite with probability 1 and  $\|E.[Z]\| < \infty$ .

If  $\alpha_1 = \alpha_2 = 1 - \theta$  and  $\alpha_3 = \alpha_4 = 1 - \theta'$ , then

$$E_{p,p'}[Z] = (1 - p)/\theta + 2(1 - p')/\theta'. \quad (3.18)$$

Proof. The first statement follows directly from Theorem 2.3. The second follows from (2.10) on taking  $\psi(p, p') = p^v p'^w$  and noting that  $m(\psi) \leq (v^2 + w^2)^{1/2}$  as a consequence of the mean value theorem and the Schwartz inequality. The third statement follows from Corollary 2.5 on taking  $A^{(1)} = \{(A, B), (\tilde{A}, B)\}$  so that  $P_{p,p'}(A^{(1)}) = p(1 - p') + (1 - p)/2$ . Since the function  $\chi(p, p') = (1 - p)/\theta + 2(1 - p')/\theta'$  is obviously continuous with  $\chi(1, 1) = 0$ , (3.18) is proved by verifying that this function satisfies the functional equation given in the statement of Corollary 2.5 when  $\alpha_1 = \alpha_2 = 1 - \theta$  and  $\alpha_3 = \alpha_4 = 1 - \theta'$ . Q.E.D.

Of the two learning rate parameters appearing in (3.18),  $\theta'$  is associated with the response learning process  $\{P_n(W|A)\}$ , while  $\theta$  is associated with the perceptual learning process  $\{P_n(A)\}$ . Suppose that the discrimination problem under consideration (with  $P_1(A) = p$  and  $P_1(W|A) = p'$ ) has been preceded by  $j$  trials of a previous (reversed) problem with black the positive stimulus. Then  $p$  will tend to increase and  $p'$  to decrease as  $j$  increases. Thus overtraining tends to decrease  $(1 - p)/\theta$  and to increase  $2(1 - p')/\theta'$ . Which effect predominates and determines the effect of overtraining on  $E_{p,p'}[Z]$  will depend on the magnitudes of  $\theta$  and  $\theta'$ , large  $\theta$  and small  $\theta'$  leading to an increase in errors with overtraining, and small  $\theta$  and large  $\theta'$  leading to a decrease in errors with overtraining -- the "overlearning reversal effect." This oversimplified argument ignores the

effect of the magnitudes of  $\theta$  and  $\theta'$  on  $p$  and  $p'$ , but it none the less suggests the power of (3.18).

In concluding this subsection it is worth remarking that the theory of Section 2 is also applicable to Bush's (1965, pp. 172-175) linear operator version of Wyckoff's (1952) discrimination model when  $P = 1/2$ . In that case,  $(x_n, y_n, u_n)$  can be taken to be the state on trial  $n$ , and this triple determines the error probability on trial  $n$ . When there are only two learning rate parameters,  $\theta' > 0$  for  $(x_n)$  and  $(y_n)$ , and  $\theta > 0$  for  $(u_n)$ , the expected total errors is given by

$$E_{x,y,u}[Z] = (1 - u)/\theta + 2(1 - x)/\theta' + 2y/\theta'.$$

## 4. SOME REMARKS ON PREVIOUS WORK

Theorem 2.2 includes the main convergence theorem of Onicescu and Mihoc (1935, Section 5), and an ergodic theorem of Ionescu Tulcea (1959, Section 8). It includes many of Karlin's (1953) results, and has points of contact with the work of Lamperti and Suppes (1959) and Iosifescu and Theodorescu (1965). None of this previous work covers the general non-cyclic four-operator (linear) model without absorbing states. Karlin's results are concerned, for the most part, with two-operator models, that is, four-operator models with  $\theta_{11} = \theta_{21} = \theta_1$  and  $\theta_{12} = \theta_{22} = \theta_2$ . The main ergodic theorem of Iosifescu and Theodorescu (1965, Theorem 2) is not applicable to any four-operator model, since one of its assumptions is that there is some positive integer  $k$ , positive real number  $\alpha$ , and response  $A_{i_0}$ , such that response  $A_{i_0}$  has probability at least  $\alpha$  on trial  $k + 1$ , regardless of the initial probability of  $A_{i_0}$  and the responses and outcomes on trials 1 through  $k$ . Such a hypothesis would be more appropriate if some of the operators in (3.3) had fixed points other than 0 or 1.

The method of Lamperti and Suppes is somewhat different from that of the present paper, and has a certain shortcoming. Consider a two-operator model with  $\theta_1, \theta_2 > 0$  and  $0 < \pi_{12}, \pi_{21} < 1$ . Such a model satisfies the hypotheses of Lamperti and Suppes' Theorem 4.1 (with  $m^* = 1$ ,  $k^* = 1$  or  $2$ , and  $m_0 = 0$ ) if their event " $E_n = j$ " is identified with " $O_{1j,n}$  or  $O_{2j,n}$ " in the notation of Subsection 3a.

One of the conclusions of that theorem is that, for all positive integers  $v$ ,  $\alpha_{1,n}^v = E_p[p_n^v]$  converges, as  $n \rightarrow \infty$ , to a quantity  $\alpha_1^v = E[p_\infty^v]$  which does not depend on the initial  $A_1$  response probability  $p$ . The  $\alpha$  notation is theirs. This conclusion follows, of course, from Theorem 2.2 of the present paper (along with an estimate of the rate of convergence that their method does not yield). But the author has found no arguments in their paper that bear directly on the lack of dependence of the limit on  $p$ . (Their notation, e.g.  $\alpha_{1,n}^v$ , does not even refer to  $p$ .) The only kind of conclusion that can be drawn from the arguments given by Lamperti and Suppes is that (in the notation of the present paper), for any  $p$ ,

$$E_p[p_{n+k}^v | O_{i_k j_k, k} A_{i_k, k} \dots O_{i_1 j_1, 1} A_{i_1, 1}] = E_{f_{i_1 j_1, \dots, i_k j_k}(p)}[p_n^v]$$

converges as  $n \rightarrow \infty$  to a quantity that does not depend on  $k$  or  $i_1, j_1, \dots, i_k, j_k$ . The recent corrections (Lamperti and Suppes (1965)) of the Lamperti and Suppes paper do not affect this observation. The method of Lamperti and Suppes is an extension of the method used

by Doeblin and Fortet (1937) to study what they call chaînes (B).

It appears that Doeblin and Fortet's treatment of Onicescu and Mihoc's chaînes à liaisons complètes (chaînes (O - M)) by means of their theory of chaînes (A) has the same shortcoming.

A distance diminishing four-operator model with two absorbing states necessarily has  $\pi_{ii}, \theta_{ii} > 0$  for  $i = 1, 2$  and either

$\pi_{ij} = 0$  or  $\theta_{ij} = 0$  for  $i \neq j$ . Thus it has two effective operators and, perhaps, an identity operator. Such models were studied by Karlin (1953), and the implications of Theorem 2.3 for these models do not add much to his results. The generality of Theorem 2.3 is roughly comparable to that of Kennedy's (1957) theorems, though Kennedy's assumptions exclude Lovejoy's Model I.

The ergodic theorem of Ionescu Tulcea and Marinescu (1950) used in the proof of Theorem 2.1 - 2.3 extends earlier work by Doeblin and Fortet (1937, see section titled "Notes sur une équation fonctionnelle"). The condition H9 was used by Jamison (1964).

Let  $Y$  be the total number of response alternations for a distance diminishing four-operator model with two absorbing barriers. That  $Y$  is finite with probability 1 (see Theorem 3.4) follows, in the special case  $\pi_{ii} = 1$ ,  $\theta_{ii} = \theta > 0$ ,  $1 \leq i \leq 2$ , from a result of Rose (1964, Corollary 2 of Theorem 5).

Theorems 3 and 4 of Iosifescu and Theodorescu (1965) give results like those of Theorem 2.6 of the present paper for a subclass of the class of models to which their Theorem 2 is applicable. This class of models is disjoint from the class of four-operator models, as was pointed out above. However, once Theorem 2.4 has been proved, a theorem of Iosifescu (1963) leads to Theorem 2.6. To the results in Theorem 3.5 and its five-operator generalization could be added

$$\frac{1}{n} \text{var}_p(S_{m,n}) = \sigma^2 + o(n^{-1/2})$$

which also follows from Theorem 2.6. In the special case of noncontingent reinforcement and  $c = 1$ , the result is a consequence of Theorem 8.10 of Estes and Suppes (1959). A similar result for  $\lim_{m \rightarrow \infty} \text{var}_p(S_{m,n})$  when reinforcement is noncontingent and  $0 < c \leq 1$  follows from formula (2.16) of Yellott (1965).

## 5. PROOFS OF THEOREM CONCERNING STATES

a. The Basic Ergodic Theorem

In this section only,  $C(S)$  is the set of complex valued continuous functions on  $S$ , and  $m(\cdot)$ ,  $|\cdot|$ ,  $\|\cdot\|$ , and  $CL$  are redefined accordingly (see (1.2), (1.3), (2.4) and the sentence following (2.4)). The spaces  $C(S)$  and  $CL$  are Banach spaces with respect to the norms  $|\cdot|$  and  $\|\cdot\|$  respectively. The space  $CL$  is also a normed linear space with respect to  $|\cdot|$ . The norm of a bounded linear operator on  $C(S)$  or  $CL$  is denoted in the same way as the norm of an element of these spaces. Thus if  $U$  is a bounded linear operator on  $C(S)$  its norm is

$$|U| = \sup_{\substack{\psi \in C(S): \\ |\psi| \leq 1}} |U\psi|,$$

while if  $U$  is a bounded linear operator on  $CL$  its norm is

$$\|U\| = \sup_{\substack{\psi \in CL: \\ \|\psi\| \leq 1}} \|U\psi\|.$$

Finally, if  $U$  is an operator on  $CL$ , bounded with respect to  $|\cdot|$ , its norm is denoted  $|U|_{CL}$ , thus

$$|U|_{CL} = \sup_{\substack{\psi \in CL: \\ |\psi| \leq 1}} |U\psi|.$$

If  $U$  is a linear operator on a linear space  $W$  over the complex

numbers, and if  $\lambda$  is a complex number,  $D(\lambda)$  denotes the set of all  $x \in W$  such that  $Ux = \lambda x$ . Obviously  $D(\lambda)$  is a linear subspace of  $W$  and always contains 0. If  $D(\lambda)$  contains an element  $x \neq 0$ ,  $\lambda$  is an eigenvalue of  $U$ .

One of the mathematical cornerstones of this paper is the following lemma, which is a specialization of a uniform ergodic theorem of Ionescu Tulcea and Marinescu (1950, Section 9) along lines suggested by these authors (Ionescu Tulcea and Marinescu, 1950, Section 10).

Lemma 5.1. Let  $U$  be a linear operator on  $CL$  such that

(i)  $|U|_{CL} \leq 1$ ,

(ii)  $U$  is bounded with respect to the norm  $\|\cdot\|$ , and

(iii) for some positive integer  $k$  and real numbers  $0 \leq r < 1$  and  $R < \infty$

$$m(U^k \psi) \leq rm(\psi) + R|\psi|$$

for all  $\psi \in CL$ . Then

(a) there are at most a finite number of eigenvalues

$\lambda_1, \lambda_2, \dots, \lambda_p$  of  $U$  for which  $|\lambda_i| = 1$ ,

(b) for all positive integers  $n$

$$U^n = \sum_{i=1}^p \lambda_i^n U_i + V^n,$$

where  $V$  and the  $U_i$  are linear operators on  $CL$ , bounded with  
respect to  $\|\cdot\|$ ,

$$(c) \quad U_i^2 = U_i, \quad U_i U_j = 0 \text{ for } i \neq j, \quad U_i V = V U_i = 0,$$

$$(d) \quad D(\lambda_i) = U_i(CL) \text{ is finite dimensional, } i = 1, \dots, p, \text{ and}$$

$$(e) \quad \text{for some } M < \infty \text{ and } h > 0,$$

$$\|V^n\| \leq M/(1+h)^n$$

for all positive integers  $n$ .

This lemma will be applied to the restriction to  $CL$  of the  
bounded linear operator

$$U\psi(s) = E_s[\psi(S_2)] = \sum_{e \in E} \psi(f_e(s)) q_e(s) \quad (5.1)$$

on  $C(S)$  associated with any distance diminishing model. This  
operator is of interest because its  $(n-1)^{\text{st}}$  iterate, applied to  
a function  $\psi \in C(S)$ , gives the expectation of  $\psi(S_n)$  as a function  
of the initial state; that is,

$$E_s[\psi(S_n)] = U^{n-1}\psi(s), \quad (5.2)$$

$n \geq 0$ . This formula is easily proved by induction. It holds by  
definition for  $n = 1$  and  $n = 2$ , and, if it holds for an  $n \geq 1$ ,  
then

$$\begin{aligned}
E_s[\psi(S_{n+1})] &= E_s[E[\psi(S_{n+1})|S_2]] \\
&= E_s[U^{n-1}\psi(S_2)] = U^n\psi(s).
\end{aligned}$$

Theorem 5.1. The conclusions (a) - (e) of Lemma 5.1 hold for the operator  $U\psi(s) = E_s[\psi(S_2)]$  associated with any distance diminishing model. In addition, 1 is an eigenvalue of  $U$  and  $D(1)$  contains all constant functions on  $S$ .

Throughout the rest of this paper it will be assumed, without loss of generality, that  $\lambda_1 = 1$  and  $\lambda_j \neq 1$ ,  $j = 2, \dots, p$ , where the  $\lambda_i$ 's are the eigenvalues of  $U$  of modulus 1.

Proof. The last statement of the theorem is obvious. It thus remains only to verify the hypotheses of Lemma 5.1. For any  $\psi \in CL$

$$\begin{aligned}
U\psi(s) - U\psi(s') &= \\
&= \sum_{\alpha} \left( \psi(f_{\alpha}(s)) - \psi(f_{\alpha}(s')) \right) \phi_{\alpha}(s) + \sum_{\alpha} \psi(f_{\alpha}(s')) (\phi_{\alpha}(s) - \phi_{\alpha}(s')),
\end{aligned}$$

so

$$\begin{aligned}
|U\psi(s) - U\psi(s')| &\leq \\
&\leq \sum_{\alpha} |\psi(f_{\alpha}(s)) - \psi(f_{\alpha}(s'))| \phi_{\alpha}(s) + \sum_{\alpha} |\psi(f_{\alpha}(s'))| |\phi_{\alpha}(s) - \phi_{\alpha}(s')| \\
&\leq m(\psi) \sum_{\alpha} d(f_{\alpha}(s), f_{\alpha}(s')) \phi_{\alpha}(s) + |\psi| \left( \sum_{\alpha} m(\phi_{\alpha}) \right) d(s, s') \\
&\leq \left[ m(\psi) \sum_{\alpha} \phi_{\alpha}(s) + |\psi| \left( \sum_{\alpha} m(\phi_{\alpha}) \right) \right] d(s, s').
\end{aligned}$$

Norman

448.

by H7. Thus  $U\psi \in CL$  as a consequence of H6, and

$$m(U\psi) \leq m(\psi) + |\psi| \sum_{\bullet} m(\varphi_{\bullet}). \quad (5.3)$$

Clearly

$$|U\psi(s)| \leq \sum_{\bullet} |\psi(f_{\bullet}(s))| \varphi_{\bullet}(s) \leq |\psi|,$$

so

$$|U\psi| \leq |\psi|.$$

Therefore (i) is satisfied, and

$$\|U\psi\| \leq m(\psi) + |\psi| (1 + \sum_{\bullet} m(\varphi_{\bullet}))$$

$$= \left(1 + \sum_{\bullet} m(\varphi_{\bullet})\right) \|\psi\|,$$

so (ii) is satisfied also.

Hypothesis (iii) of Lemma 5.1 will now be verified. Let  $k(s)$  be the integer and  $e_{s,1} \dots e_{s,k(s)}$  the events in the statement of H8. For each  $s \in S$  let

$$Z(s) = \{t \in S: \varphi_{e_{s,1} \dots e_{s,k(s)}}(t) > 0\}.$$

Since

$$\varphi_{e_1 \dots e_n}(t) = \varphi_{e_1}(t) \varphi_{e_2}(f_{e_1}(t)) \dots \varphi_{e_n}(f_{e_1 \dots e_{n-1}}(t))$$

is continuous for any events  $e_1, \dots, e_n$ ,  $Z(s)$  is open. Furthermore  $s \in Z(s)$ . Since  $S$  is compact the open covering  $\{Z(s): s \in S\}$  has a finite subcovering  $Z(s_1), \dots, Z(s_m)$ . Let  $k_1 = k(s_1)$  and let  $K = \max_{1 \leq i \leq m} k_i$ . If  $t \in Z(s_1)$ , let  $e'_{t,j} = e_{s_1,j}$ ,  $j = 1, \dots, k_1$ . Clearly it is possible to choose  $e'_{t,k_1+1}, \dots, e'_{t,K}$  in such a way that

$$\varphi_{e'_{t,1} \dots e'_{t,K}}(t) > 0.$$

Hypothesis H7 implies  $\mu(f_{e_1 \dots e_j}) \leq 1$  for any events  $e_1, \dots, e_j$ , so the inequality

$$\mu(f_{e'_{t,1} \dots e'_{t,K}}) < 1$$

is obtained from H8. Thus the integer  $K$ , which does not depend on  $t$ , and the system  $e'_{t,1}, \dots, e'_{t,K}$  of events satisfy H8. Therefore it can be assumed without loss of generality that the integer  $k$  in H8 does not depend on the state.

If  $\psi \in CL$  and  $s, s' \in S$  then

$$\begin{aligned} U^k \psi(s) - U^k \psi(s') &= \\ &= \sum_{e_1 \dots e_k} \left( \psi(f_{e_1 \dots e_k}(s)) - \psi(f_{e_1 \dots e_k}(s')) \right) \varphi_{e_1 \dots e_k}(s) \\ &\quad + \sum_{e_1 \dots e_k} \psi(f_{e_1 \dots e_k}(s')) (\varphi_{e_1 \dots e_k}(s) - \varphi_{e_1 \dots e_k}(s')). \end{aligned}$$

Therefore

$$\begin{aligned} |U^k \psi(s) - U^k \psi(s')| &\leq \\ &= \left[ m(\psi) - \sum_{e_1 \dots e_k} \mu(f_{e_1 \dots e_k}) \varphi_{e_1 \dots e_k}(s) + |\psi| \sum_{e_1 \dots e_k} m(\varphi_{e_1 \dots e_k}) \right] d(s, s'). \end{aligned}$$

Now  $CL$  is closed under multiplication and under composition with mappings of  $S$  into  $S$  for which  $\mu(f) < \infty$ . Thus  $\varphi_{e_1 \dots e_j} \in CL$  for any events  $e_1, \dots, e_j$  and

$$m_j = \sum_{e_1 \dots e_j} m(\varphi_{e_1 \dots e_j}) < \infty. \quad (5.4)$$

It follows that for  $s \neq s'$

$$\begin{aligned}
 & |U^k \psi(s) - U^k \psi(s')| / d(s, s') \leq \\
 & m(\psi) \left[ \sum_{\substack{e_1 \dots e_k: \\ \mu(f_{e_1 \dots e_k}) < 1}} + \sum_{\substack{e_1 \dots e_k: \\ \mu(f_{e_1 \dots e_k}) = 1}} \right] \mu(f_{e_1 \dots e_k}) \varphi_{e_1 \dots e_k}(s) + |\psi|_{m_k} \\
 & \leq m(\psi) \left[ \lambda \sum_{\substack{e_1 \dots e_k: \\ \mu(f_{e_1 \dots e_k}) < 1}} + \sum_{\substack{e_1 \dots e_k: \\ \mu(f_{e_1 \dots e_k}) = 1}} \right] \varphi_{e_1 \dots e_k}(s) + |\psi|_{m_k} \\
 & \leq m(\psi) (\lambda \Delta + (1 - \Delta)) + |\psi|_{m_k},
 \end{aligned}$$

where

$$\lambda = \max_{\substack{e_1 \dots e_k: \\ \mu(f_{e_1 \dots e_k}) < 1}} \mu(f_{e_1 \dots e_k}) < 1 \quad (5.5)$$

and

$$\Delta = \min_{s \in S} \sum_{\substack{e_1 \dots e_k: \\ \mu(f_{e_1 \dots e_k}) < 1}} \varphi_{e_1 \dots e_k}(s) > 0. \quad (5.6)$$

The latter inequality is a consequence of H3, H5, and the continuity of the  $\varphi_{e_1 \dots e_k}$ . Therefore

$$m(U^k \psi) \leq rm(\psi) + m_k |\psi| \quad (5.7)$$

where  $r = \lambda\Delta + (1 - \Delta) < 1$ , and (iii) of Lemma 5.1 is satisfied. Q.E.D.

Though it constitutes a slight digression and will not be used in the sequel, an additional consequence of (5.3) and (5.7) is worth pointing out. From these inequalities

$$m(U^n \psi) \leq m(\psi) + n|\psi|m_1 \quad \text{and} \quad (5.8)$$

$$m(U^{nk} \psi) \leq m(\psi) + |\psi|m_k/(1 - r), \quad (5.9)$$

$n \geq 0$ , are easily obtained by induction. These yield, on combination,

$$m(U^j \psi) \leq m(\psi) + |\psi|((k - 1)m_1 + m_k/(1 - r)), \quad (5.10)$$

valid for all  $\psi \in CL$  and  $j \geq 0$ . It follows that  $\{U^j \psi\}$  is equicontinuous. This, together with the fact that  $CL$  is dense in  $C(S)$  (as a consequence of the Stone-Weierstrass Theorem) and  $|U| = 1$ , implies that, for any  $\psi \in C(S)$ ,  $\{U^j \psi\}$  is equicontinuous. In the terminology of Jamison (1964, <sup>1965</sup>) the operator  $U$  on  $C(S)$  associated with any distance diminishing model is uniformly stable. In the terminology of Feller (1966) the corresponding stochastic kernel  $K$  is regular.

#### b. Proof of Theorem 2.1

The following lemma includes most of the assertions of Theorem 2.1.

Lemma 5.4 For any distance diminishing model there is a stochastic kernel  $K^n$  such that (2.5) holds, where  $\pi_n[v(s_n)]$  is given by (2.6).

Proof of Lemma 5.2. Theorem 5.1 implies that

$$\|u^n\| \leq \sum_{i=1}^p \|u_i\| + \|v^n\| \rightarrow \sum_{i=1}^p \|u_i\|$$

as  $n \rightarrow \infty$ . Therefore there is a constant  $W < \infty$  such that

$$\|u^n\| < W \quad (5.11)$$

for all  $n \geq 0$ .

Let  $\bar{u}_n = (1/n) \sum_{j=0}^{n-1} u^j$ . Then, by Theorem 5.1,

$$\begin{aligned} \bar{u}_n &= (1/n)(I - u^n) + (1/n) \sum_{j=1}^n u^j \\ &= (1/n)(I - u^n) + (1/n) \sum_{i=1}^p \left( \sum_{j=1}^n \lambda_i^j \right) u_i + (1/n) \sum_{j=1}^n v^j. \end{aligned}$$

Therefore

$$\bar{u}_n - u_1 = (1/n)(I - u^n) + (1/n) \sum_{i=2}^p \lambda_i \left[ (1 - \lambda_i^n) / (1 - \lambda_i) \right] u_i + (1/n) \sum_{j=1}^n v^j,$$

so that

$$\|\bar{u}_n - u_1\| \leq c/n$$

where

$$C = (1 + W) + 2 \sum_{i=2}^P \|U_i\|/|1 - \lambda_i| + M/h.$$

Thus, for any  $\psi \in CL$ ,

$$\|\bar{U}_n \psi - U_1 \psi\| \leq \|\bar{U}_n - U_1\| \|\psi\| \leq C \|\psi\|/h, \quad (5.12)$$

and, a fortiori,

$$\lim_{n \rightarrow \infty} \|\bar{U}_n \psi - U_1 \psi\| = 0 \quad (5.13)$$

for all  $\psi \in CL$ .

Since  $CL$  is dense in  $C(S)$  and  $|\bar{U}_n| = 1$  for  $n \geq 1$ , it follows that (5.13) holds for all  $\psi \in C(S)$ , where  $U_1$  has been extended (uniquely) to a bounded linear operator on  $C(S)$ . Since the operators  $\bar{U}_n$  on  $C(S)$  are all positive and preserve constants, (5.13) implies that the same is true of  $U_1$ . Thus, for any  $s \in S$ ,  $U_1 \psi(s)$  is a positive linear functional on  $C(S)$  with  $U_1 \Gamma(s) = 1$  where  $\Gamma(s) = 1$ . Hence, by the Riesz representation theorem, there is a (unique) Borel probability measure  $K^n(s, \cdot)$  on  $S$  such that

$$U_1 \psi(s) = \int_S \psi(s') K^n(s, ds') \quad (5.14)$$

for all  $\psi \in C(S)$ . In view of (5.14), (5.12) reduces to (2.5).

That  $K^\pi$  is a stochastic kernel follows from the fact, now to be proved, that  $K^\pi(\cdot, A) \in CL$  for every Borel subset  $A$ . This is obviously true if  $A = S$ . Suppose that  $A$  is an open set such that its complement  $\tilde{A}$  is not empty. For  $j \geq 1$  define  $\eta_j \in CL$  by

$$\eta_j(s) = \begin{cases} 1 & \text{if } d(s, \tilde{A}) \geq 1/j \\ jd(s, \tilde{A}) & \text{if } d(s, \tilde{A}) \leq 1/j. \end{cases} \quad (5.15)$$

Then

$$\lim_{j \rightarrow \infty} \eta_j(s) = \begin{cases} 1 & \text{if } s \in A \\ 0 & \text{if } s \in \tilde{A} \end{cases} = I_A, \quad (5.16)$$

where  $I_A$  is the indicator function of the set  $A$ , and the convergence is monotonic. Therefore

$$\lim_{j \rightarrow \infty} U_1 \eta_j(s) = \int_S I_A(s') K^\pi(s, ds') = K^\pi(s, A) \quad (5.17)$$

for all  $s \in S$ . By Theorem 5.1,  $D(1) = U_1(CL)$  is a finite dimensional subspace of  $CL$ . Hence there exists a constant  $J < \infty$  such that  $\|\psi\| \leq J|\psi|$  for all  $\psi \in D(1)$ . Therefore

$$|U_1 \eta_j(s_1) - U_1 \eta_j(s_2)| \leq \pi(U_1 \eta_j) d(s_1, s_2)$$

$$\leq J|U_1 \eta_j| d(s_1, s_2) \leq Jd(s_1, s_2)$$

for all  $j \geq 1$  and  $s_1, s_2 \in S$ . Equation (5.17) then yields, on letting  $j$  approach  $\infty$ ,

$$|K^\pi(s_1, A) - K^\pi(s_2, A)| \leq Jd(s_1, s_2).$$

If  $A$  is an arbitrary Borel set,  $s_1, s_2 \in S$ , and  $\epsilon > 0$ , the regularity of  $K^\omega(s_1, \cdot)$  insures the existence of an open set  $A_{1,\epsilon}$  such that  $A_{1,\epsilon} \supset A$  and

$$K^\omega(s_1, A_{1,\epsilon}) - K^\omega(s_1, A) \leq \epsilon$$

for  $i = 1, 2$ . Thus  $A_\epsilon = A_{1,\epsilon} \cap A_{2,\epsilon}$  is open and

$$0 \leq K^\omega(s_1, A_\epsilon) - K^\omega(s_1, A) \leq \epsilon,$$

$i = 1, 2$ . Combination of these inequalities with the result of the last paragraph yields

$$|K^\omega(s_1, A) - K^\omega(s_2, A)| \leq Jd(s_1, s_2) + 2\epsilon,$$

or, since  $\epsilon$  is arbitrary,

$$|K^\omega(s_1, A) - K^\omega(s_2, A)| \leq Jd(s_1, s_2).$$

Thus  $K^\omega(\cdot, A) \in CL$  with  $m(K^\omega(\cdot, A)) \leq J$  for all Borel subsets  $A$  of  $S$ . Q.E.D.

Actually, this proof gives (2.5) for the complex as well as real valued functions  $\psi$ , though this is not important here.

To complete the proof of Theorem 2.1 it remains only to prove

**Lemma 5.3.** The stochastic kernel  $(1/n)\sum_{j=0}^{n-1} K^{(j)}$  converges uniformly to  $K^\omega$ .

Proof. Denote  $(1/n)\sum_{j=0}^{n-1} K^{(j)}$  by  $\bar{K}_n$ .

Since

$$\bar{K}_n(s, \bar{G}) \leq \bar{K}_n(s, G) \leq \bar{K}_n(s, \bar{G}) \quad (5.18)$$

it suffices to show that if  $A$  is open,

$$K^n(s, A) - \epsilon \leq \bar{K}_n(s, A) \quad (5.19)$$

for all  $s$  if  $n$  is sufficiently large, while if  $B$  is closed,

$$\bar{K}_n(s, B) \leq K^n(s, B) + \epsilon$$

for all  $s$  if  $n$  is sufficiently large. The statement concerning closed sets follows from that concerning open sets by taking complements, so only open sets need be considered. There is no loss in generality in assuming  $A \neq \emptyset$ . By (5.15),  $I_A(t) = \eta_j(t)$  for all  $t \in S$  so

$$\begin{aligned} \bar{K}_n(s, A) &= \bar{U}_n \eta_j(s) \\ &= K^n(s, A) + [U_1 \eta_j(s) - K^n(s, A)] + [\bar{U}_n \eta_j(s) - U_1 \eta_j(s)] \\ &= K^n(s, A) - |U_1 \eta_j(\cdot) - K^n(\cdot, A)| - |\bar{U}_n \eta_j - U_1 \eta_j|. \end{aligned}$$

Since the convergence in (5.17) is monotonic and the limit is continuous, convergence is uniform by Dini's theorem. Choose  $j$  so large that

$$|U_1 \eta_j(\cdot) - K^n(\cdot, A)| < \epsilon/2.$$

Then (5.13) applied to  $\psi = \eta_j$  implies (5.19) for all  $s \in S$  if  $n$  is sufficiently large. Q.E.D.

Theorem 5.1 asserts that  $D(1)$ , the linear space of  $\psi \in CL$  such that  $U\psi = \psi$ , contains all constant functions. If, in addition, it is known that  $D(1)$  is one dimensional; that is, the only  $\psi \in CL$  such that  $U\psi = \psi$  are constants, then it can be concluded that the probability measure  $K^\pi(s, \cdot) = K^\pi(\cdot)$  does not depend on  $s$ . For, by (d) of Lemma 5.1,  $U_1\psi$  is a constant function for any  $\psi \in CL$ , and thus for any  $\psi \in C(S)$ . Therefore, in view of (5.14), for any  $s, s' \in S$ ,

$$\int_S \psi(t) K^\pi(s, dt) = \int_S \psi(t) K^\pi(s', dt)$$

for all  $\psi \in C(S)$ . This implies that  $K^\pi(s, \cdot) = K^\pi(s', \cdot)$ , as claimed. It is, incidentally, easy to show that  $K^\pi(\cdot)$  is the unique stationary probability distribution of  $\{s_n\}$ , from which it follows (Breiman (1960)) that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{j=1}^n \psi(s_j) = E[\psi(s_\infty)]$$

with probability 1, for any  $\psi \in C(S)$  and any initial state.

c. Proof of Theorem 2.2

Suppose that a distance diminishing model has the property that the associated operator  $U$  on  $CL$  has no eigenvalues of modulus 1 other than 1. Then Theorem 5.1 implies that

$$\|U^n - U_1\| = \|v^n\| \leq M/(1+h)^n.$$

Therefore, for any  $\psi \in CL$

$$\|U^n \psi - U_1 \psi\| \leq M\|\psi\|/(1+h)^n, \quad (5.20)$$

and, for any  $\psi \in C(S)$ ,

$$\lim_{n \rightarrow \infty} |U^n \psi - U_1 \psi| = 0. \quad (5.21)$$

From (5.21) it follows that  $K^{(n)}$  converges uniformly to  $K^\infty$ , just as uniform convergence of  $\bar{K}_n$  to  $K^\infty$  followed from (5.13) in Subsection 5b. That is to say, when (5.21) holds the proof of Lemma 5.3 remains valid if  $\bar{K}_n$  and  $\bar{U}_n$  are everywhere replaced by  $K^{(n)}$  and  $U^n$ . If, in addition, the only  $\psi \in CL$  for which  $U\psi = \psi$  are constants, then  $K^\infty(s, \cdot) = K^\infty(\cdot)$  does not depend on  $s$ , as was shown in the last paragraph of Subsection 5b. Therefore (5.20) reduces to (2.9) (with  $C = (1+h)M$  and  $\alpha = 1/(1+h)$ ), and all of the conclusions of Theorem 2.2 hold. To complete the proof of Theorem 2.2 it thus suffices to prove the following two lemmas.

Lemma 5.4. If a distance diminishing model satisfies H9,  
then 1 is the only eigenvalue of U of modulus 1.

Lemma 5.5. If a distance diminishing model satisfies H9,  
then the only continuous solutions of  $U\psi = \psi$  are constants.

The arguments given below follow similar arguments by Jamison (1964).

Proof of Lemma 5.4. Suppose  $|\lambda| = 1$ ,  $\lambda \neq 1$ ,  $\psi \neq 0$ , and  $\psi \in D(\lambda)$  so that  $U\psi = \lambda\psi$ .  $|\psi(\cdot)| \in C(S)$  so there is an  $s_0 \in S$ :

$$|\psi(s_0)| = \max_{s \in S} |\psi(s)| = |\psi|.$$

Clearly  $\psi(s_0) \neq 0$ . Now  $U\psi' = \lambda\psi'$  where  $\psi' = \psi/\psi(s_0)$ ,  $|\psi'| = |\psi|/|\psi(s_0)| = 1$  and  $\psi'(s_0) = 1$ .  $U^n\psi' = \lambda^n\psi'$  so  $U^n\psi'(s_0) = \lambda^n$ . For  $n = 1, 2, \dots$  let  $B_n = \{s: \psi'(s) = \lambda^n\}$ . Clearly  $K^{(n)}(s_0, B_n) = 1$  for all  $n \geq 1$ . Since  $K(s_0, B_1) = 1$ ,  $B_1$  is not empty. Let  $s_1 \in B_1$ . Then  $\psi'(s_1) = \lambda$ ,  $U^n\psi'(s_1) = \lambda^{n+1}$ , and  $K^{(n)}(s_1, B_{n+1}) = 1$ . But  $|\lambda^{n+1} - \lambda^n| = |\lambda - 1| > 0$ . Since  $\psi'$  is uniformly continuous there exists  $\delta > 0$  such that  $d(s', s'') < \delta$  implies  $|\psi'(s') - \psi'(s'')| < |\lambda - 1|$ . If  $s' \in B_n$  and  $s'' \in B_{n+1}$  then  $|\psi'(s') - \psi'(s'')| = |\lambda - 1|$  so  $d(s', s'') \geq \delta$ . Therefore  $d(B_n, B_{n+1}) \geq \delta$ ,  $n = 1, 2, \dots$ . But  $B_n \supset T_n(s_0)$  and  $B_{n+1} \supset T_n(s_1)$  so  $d(T_n(s_0), T_n(s_1)) \geq \delta$ ,  $n = 1, 2, \dots$ . Thus H9 is violated. Q.E.D.

Proof of Lemma 5.5. Suppose that there exists a real valued non-constant function  $\psi \in C(S)$  such that  $U\psi = \psi$ . Let  $M = \max \psi = \psi(s_0)$  and  $m = \min \psi = \psi(s_1)$ . Then  $M > m$ . Let  $C_m = \{s: \psi(s) = m\}$  and  $C_M = \{s: \psi(s) = M\}$ .  $U^n \psi = \psi$ , so  $U^n \psi(s_0) = M$  and  $U^n \psi(s_1) = m$ . Therefore  $K^{(n)}(s_0, C_M) = 1$  and  $K^{(n)}(s_1, C_m) = 1$ , so  $C_M \supset T_n(s_0)$  and  $C_m \supset T_n(s_1)$  for all  $n \geq 1$ . By the uniform continuity of  $\psi$  there exists a  $\delta > 0$  such that  $d(s, s') < \delta$  implies  $|\psi(s) - \psi(s')| < M - m$ . If  $s \in C_M$  and  $s' \in C_m$  then  $|\psi(s) - \psi(s')| = M - m$ , so  $d(s, s') \geq \delta$ . Therefore  $d(T_n(s_0), T_n(s_1)) \geq \delta$  for all  $n \geq 1$ . Thus H9 is violated. So under the hypotheses of Lemma 5.5 there is no real valued non-constant  $\psi \in C(S)$  for which  $U\psi = \psi$ .

Suppose  $\psi' \in C(S)$ ,  $U\psi' = \psi'$ . Then

$$U \operatorname{Re} \psi' + iU \operatorname{Im} \psi' = \operatorname{Re} \psi' + i \operatorname{Im} \psi'.$$

Thus  $U \operatorname{Re} \psi' = \operatorname{Re} \psi'$  and  $U \operatorname{Im} \psi' = \operatorname{Im} \psi'$ . But  $\operatorname{Re} \psi'$  and  $\operatorname{Im} \psi'$  are continuous and real valued, so  $\operatorname{Re} \psi'$  and  $\operatorname{Im} \psi'$  are constants.

Thus  $\psi'$  is a (complex valued) constant function. Therefore all continuous solutions of  $U\psi = \psi$  are constants. Q.E.D.

#### d. Proof of Theorem 2.3

The first paragraph of Subsection 5c shows that, to obtain the uniform convergence of  $r^{(n)}$  to the limiting kernel  $K^\infty$  of Theorem 2.1, and to obtain (2.10) with  $E_\mu[\psi(s_\mu)]$  defined as in (2.6), it suffices to prove Lemma 5.6 below. All lemmas in this subsection refer to

a distance diminishing model satisfying H10.

Lemma 5.6.  $U$  has no eigenvalues of modulus 1 other than 1.

Proof. Suppose  $U\psi = \lambda\psi$  where  $|\lambda| = 1$ ,  $\lambda \neq 1$ , and  $\psi \in C(S)$ . Let  $s_0$  be a state for which  $|\psi(s_0)| = |\psi|$ , and let  $C_n = \{s: \psi(s) = \lambda^n \psi(s_0)\}$ ,  $n = 1, 2, \dots$ . Now  $U^n \psi(s_0) = \lambda^n \psi(s_0)$ , thus  $K^{(n)}(s_0, C_n) = 1$ , and  $C_n \supset T_n(s_0)$ . By H10 there exists a sequence  $\{t_n\}$  such that  $t_n \in T_n(s_0)$  and  $\lim_{n \rightarrow \infty} d(t_n, s_j(s_0)) = 0$ . Hence  $\lim_{n \rightarrow \infty} \psi(t_n) = \psi(s_j(s_0))$ . But  $t_n \in C_n$ , so  $\psi(t_n) = \lambda^n \psi(s_0)$ , which converges only if  $\psi(s_0) = 0$ . Hence  $|\psi| = 0$  and  $\psi(s) = 0$ . Thus  $\lambda$  is not an eigenvalue of  $U$ . Q.E.D.

The proof that  $S_n$  converges with probability 1 must be deferred until more information has been obtained about  $K^\infty$ . The next two lemmas provide such information. In the work that follows,  $A = \{s_i: 1 \leq i \leq N\}$  is the set of absorbing states.

Lemma 5.7. If  $b_1, \dots, b_N$  are any  $N$  scalars, there is one and only one  $\psi \in C(S)$  such that  $U\psi = \psi$  and  $\psi(s_i) = b_i$ ,  $i = 1, \dots, N$ . This function belongs to CL.

Proof. (1) Uniqueness. First, the following maximum modulus principle will be proved: If  $\psi \in C(S)$  and  $U\psi = \psi$ , then all maxima of  $|\psi(\cdot)|$  occur on  $A$  (and possibly elsewhere). Let  $s_0$  be a state such that  $|\psi(s_0)| = |\psi|$ , and let  $C = \{s: \psi(s) = \psi(s_0)\}$ . Since  $U^n \psi(s_0) = \psi(s_0)$ ,  $K^{(n)}(s_0, C) = 1$ , so  $C \supset T_n(s_0)$ . By H10 there exists a sequence  $t_n$  such that  $t_n \in T_n(s_0)$  and  $\lim_{n \rightarrow \infty} d(t_n, s_j(s_0)) = 0$ .

Hence  $\lim_{n \rightarrow \infty} \psi(t_n) = \psi(a_j(s_0))$ . But  $t_n \in C$ , so  $\psi(t_n) = \psi(s_0)$ .

Thus  $\psi(s_0) = \psi(a_j(s_0))$ , and  $|\psi(a_j(s_0))| = |\psi|$ .

Suppose now that  $\psi, \psi' \in C(S)$ ,  $U\psi = \psi$ ,  $U\psi' = \psi'$ , and  $\psi(s) = \psi'(s)$  for all  $s \in A$ . Let  $\psi'' = \psi - \psi'$ . Then  $\psi'' \in C(S)$ ,  $U\psi'' = \psi''$ , and  $\psi''(s) = 0$  for all  $s \in A$ . Thus  $|\psi''| = 0$ , so  $\psi(s) = \psi'(s)$ .

(2) Existence. Since  $U^n \psi(s) = E_s[\psi(s_{n+1})]$ ,  $U^n \psi(s) = \psi(s)$  for all  $s \in A$  and  $\psi \in C(S)$ . Thus  $U_1 \psi(s) = \lim_{n \rightarrow \infty} U^n \psi(s) = \psi(s)$  for all  $s \in A$ .

Let  $\omega_1, \dots, \omega_N \in CL$  with  $\omega_i(a_j) = \delta_{ij}$ , e.g.,

$$\omega_i(s) = (1 - \epsilon^{-1} d(s, a_i))^+$$

where  $\epsilon = \min_{i \neq j} d(a_i, a_j)$  and  $x^+$  is  $x$  or  $0$  depending on whether  $x \geq 0$  or  $\leq 0$ . It will now be shown that

$$\gamma(s) = \sum_{i=1}^N b_i U_1 \omega_i(s)$$

is the function sought. Clearly  $\gamma \in CL$ , and

$$\gamma(a_j) = \sum_{i=1}^N b_i U_1 \omega_i(a_j) = \sum_{i=1}^N b_i \omega_i(a_j) = b_j.$$

Finally,

$$U\gamma = \sum_{i=1}^N b_i U U_1 \omega_i = \sum_{i=1}^N b_i U_1 \omega_i = \gamma. \quad \text{Q.E.D.}$$

**Lemma 5.8.** For  $i = 1, \dots, N$ , let  $\gamma_i$  be the continuous function

such that  $\gamma_i(a_j) = \delta_{ij}$  and  $U\gamma_i = \gamma_i$ . Then

$$K^\infty(s, \cdot) = \sum_{i=1}^N \gamma_i(s) \delta_{a_i}(\cdot),$$

where  $\delta_{a_i}$  is the Borel probability measure on  $S$  concentrated at  $a_i$ ,

and, for any  $\psi \in C(S)$  and all  $s \in S$ ,

$$E_s[\psi(S_\infty)] = \sum_{i=1}^N \gamma_i(s) \psi(a_i).$$

Proof. For  $\psi \in C(S)$  let  $\bar{\psi}(s) = \sum_{i=1}^N \gamma_i(s) \psi(a_i)$  and  $\psi' = U_1 \psi$ .

Clearly  $\bar{\psi}, \psi' \in C(S)$ , and  $\psi'(a_j) = \psi(a_j) = \bar{\psi}(a_j)$ ,  $j = 1, \dots, N$ .

Also

$$U\bar{\psi} = \sum_{i=1}^N \psi(a_i) U\gamma_i = \sum_{i=1}^N \psi(a_i) \gamma_i = \bar{\psi}, \text{ and}$$

$$U\psi' = UU_1 \psi = U_1 \psi = \psi'.$$

Thus, by Lemma 5.7,  $\bar{\psi} = \psi'$ , which is the second assertion of

Lemma 5.8.

Now

$$\begin{aligned} \int_S \psi(s') K^\infty(s, ds') &= E_s[\psi(S_\infty)] = \sum_{i=1}^N \gamma_i(s) \psi(a_i) \\ &= \sum_{i=1}^N \gamma_i(s) \int_S \psi(t) \delta_{a_i}(dt) \\ &= \int_S \psi(t) \left( \sum_{i=1}^N \gamma_i(s) \delta_{a_i} \right) (dt) \end{aligned}$$

for all  $\psi \in C(S)$ . This yields the first assertion of the lemma.

Q.E.D.

Now that it is known that  $K^n(s, \cdot)$  is concentrated on  $A$  and the functions  $\gamma_i$  are available, probability 1 convergence of  $S_n$  can be proved.

**Lemma 5.9.** For any initial state  $s$ ,  $(S_n)$  converges with probability 1 to a random point  $S_\infty$  of  $A$ . For any Borel subset  $B$  of  $S$ ,  $K^n(s, B) = P_s(S_\infty \in B)$ . In particular,

$$\gamma_i(s) = P_s(S_\infty = a_i),$$

$i = 1, \dots, N$ .

**Proof.** It is a simple consequence of the triangle inequality that the function  $d(\cdot, A)$  on  $S$  belongs to CL, and clearly  $d(a_i, A) = 0$ ,  $i = 1, \dots, N$ . Thus  $E_s[d(S_n, A)] = 0$ , so that

$$\|E_s[d(S_n, A)]\| \leq C\alpha^n \|d(\cdot, A)\|$$

for all  $n \geq 1$ . The initial state is regarded as fixed throughout the following discussion. Since

$$E_s[d(S_n, A)] \leq C\alpha^n \|d(\cdot, A)\|$$

it follows that

$$\sum_{n=1}^{\infty} E_s[d(S_n, A)] \leq C \|d(\cdot, A)\| \alpha / (1 - \alpha).$$

By the monotone convergence theorem the order of summation and expectation on the left can be interchanged to obtain

$$E_S \left[ \sum_{n=1}^{\infty} d(S_n, A) \right] < \infty.$$

Therefore  $\sum_{n=1}^{\infty} d(S_n, A) < \infty$ , and, consequently,  $\lim_{n \rightarrow \infty} d(S_n, A) = 0$  with probability 1.

For any  $i = 1, \dots, N$ ,

$$E_S [\gamma_i(S_{n+1}) | S_n, \dots, S_1] = E_S [\gamma_i(S_{n+1}) | S_n] = \gamma_i(S_n),$$

so  $\{\gamma_i(S_n)\}$  is a martingale. Since it is bounded (by  $|\gamma_i|$ ), it converges with probability 1.

Let  $G$  be the event " $\lim_{n \rightarrow \infty} d(S_n, A) = 0$  and  $\lim_{n \rightarrow \infty} \gamma_i(S_n)$  exists,  $i = 1, \dots, N$ " in the underlying sample space. The above arguments show that  $P_S(G) = 1$ . Let  $\omega \in G$ . Since  $S$  is compact, every subsequence of  $\{S_n(\omega)\}$  has a convergent subsequence, and, since  $d(S_n(\omega), A) \rightarrow 0$  as  $n \rightarrow \infty$ , all subsequential limit points of  $\{S_n(\omega)\}$  are in  $A$ . Suppose that  $a_i$  and  $a_{i'}$ ,  $i \neq i'$ , are two distinct subsequential limit points -- say  $S_{n_j}(\omega) \rightarrow a_i$  and  $S_{n'_j}(\omega) \rightarrow a_{i'}$ , as  $j \rightarrow \infty$ . Then

$$\gamma_i(S_{n_j}(\omega)) \rightarrow \gamma_i(a_i) = 1 \quad \text{and}$$

$$\gamma_{i'}(S_{n'_j}(\omega)) \rightarrow \gamma_{i'}(a_{i'}) = 0,$$

which contradicts the convergence of  $\{\gamma_i(S_n(\omega))\}$ . Thus all convergent subsequences of  $\{S_n(\omega)\}$  converge to the same point of  $A$ . Denote this point  $S_\infty(\omega)$ . It follows that  $\lim_{n \rightarrow \infty} S_n(\omega) = S_\infty(\omega)$ . Therefore  $\lim_{n \rightarrow \infty} S_n = S_\infty$  with probability 1. This implies that the asymptotic distribution of  $S_n$  is the same as the distribution of  $S_\infty$ , i.e.,  $K^\infty(s, B) = P_\mathbf{g}(S_\infty \in B)$  for all Borel subsets  $B$  of  $S$ . Finally  $\gamma_i(s) = P_\mathbf{g}(S_\infty = a_i)$  follows by taking  $B = \{a_i\}$ . Q.E.D.

This completes the proof of Theorem 2.3.

## 6. PROOFS OF THEOREMS CONCERNING EVENTS

a. Proof of Theorem 2.4

The equality

$$P_s^{(n)}(A^L) = E_s[P((E_n, \dots, E_{n+L-1}) \in A^L | S_n)]$$

can be rewritten in the form

$$P_s^{(n)}(A^L) = E_s[\psi(S_n)], \quad (6.1)$$

where

$$\psi(s) = P_s^{(1)}(A^L). \quad (6.2)$$

Thus (2.12), with  $L = C(D+1)$ , follows from (2.9), (2.10), and the following lemma.

Lemma 6.1. For any distance diminishing model there is a constant D such that

$$m(P_s^{(1)}(A^L)) \leq D \quad (6.3)$$

for all  $L \geq 1$  and  $A^L \subset E^L$ .

Proof. For any  $i, j \geq 1$ ,  $s, s' \in S$ , and  $A^{i+j} \in E^{i+j}$

$$P_s^{(1)}(A^{i+j}) - P_{s'}^{(1)}(A^{i+j}) =$$

$$= \sum_{e_1 \dots e_i} \varphi_{e_1 \dots e_i}(s) P_{f_{e_1 \dots e_i}}^{(1)}(s) (\Lambda_{e_1 \dots e_i}^{1+j}) -$$

$$\sum_{e_1 \dots e_i} \varphi_{e_1 \dots e_i}(s') P_{f_{e_1 \dots e_i}}^{(1)}(s') (\Lambda_{e_1 \dots e_i}^{1+j}),$$

where  $\Lambda_{e_1 \dots e_i}^{1+j} = \{(e_{i+1}, \dots, e_{i+j}) : (e_1, \dots, e_{i+j}) \in \Lambda^{1+j}\},$

$$= \sum_{e_1 \dots e_i} \varphi_{e_1 \dots e_i}(s) (P_{f_{e_1 \dots e_i}}^{(1)}(s) (\Lambda_{e_1 \dots e_i}^{1+j}) - P_{f_{e_1 \dots e_i}}^{(1)}(s') (\Lambda_{e_1 \dots e_i}^{1+j}))$$

$$+ \sum_{e_1 \dots e_i} (\varphi_{e_1 \dots e_i}(s) - \varphi_{e_1 \dots e_i}(s')) P_{f_{e_1 \dots e_i}}^{(1)}(s') (\Lambda_{e_1 \dots e_i}^{1+j}).$$

Thus

$$|P_s^{(1)}(\Lambda^{1+j}) - P_{s'}^{(1)}(\Lambda^{1+j})| \leq$$

$$\sum_{e_1 \dots e_i} \varphi_{e_1 \dots e_i}(s) |P_{f_{e_1 \dots e_i}}^{(1)}(s) (\Lambda_{e_1 \dots e_i}^{1+j}) - P_{f_{e_1 \dots e_i}}^{(1)}(s') (\Lambda_{e_1 \dots e_i}^{1+j})|$$

$$+ \sum_{e_1 \dots e_i} |\varphi_{e_1 \dots e_i}(s) - \varphi_{e_1 \dots e_i}(s')| P_{f_{e_1 \dots e_i}}^{(1)}(s') (\Lambda_{e_1 \dots e_i}^{1+j})$$

$$\leq n_j \sum_{e_1 \dots e_i} \varphi_{e_1 \dots e_i}(s) \mu(f_{e_1 \dots e_i}) d(s, s') + m_1 d(s, s'),$$

where

$$n_j = \max_{A^j \in \mathcal{B}^j} m(P_{\cdot}^{(1)}(A^j))$$

and  $m_1$  is given by (5.4). (Note that  $n_1 \leq m_1$ .)

Two cases are now distinguished.

Case 1:  $i = 1$ . Then

$$|P_s^{(1)}(A^{1+j}) - P_{s'}^{(1)}(A^{1+j})| \leq (n_j + m_1) d(s, s'),$$

so  $n_{j+1} \leq n_j + m_1$  or, by induction,

$$n_j \leq j m_1. \quad (6.4)$$

Case 2:  $i = k$ , where  $k$  is an integer that satisfies H8 for all  $s$ .

It was shown in the proof of Theorem 5.1 that such an integer exists, and that there is a constant  $0 \leq r < 1$  such that

$$\sum_{e_1 \dots e_k} \varphi_{e_1 \dots e_k}(s) \mu(f_{e_1 \dots e_k}) \leq r$$

for all  $s \in S$ . Thus

$$|p_s^{(1)}(\lambda^{k+j}) - p_{s'}^{(1)}(\lambda^{k+j})| \leq (n_j r + m_k) d(s, s'),$$

so

$$n_{j+k} \leq n_j r + m_k. \quad (6.5)$$

This formula and a simple induction on  $v$  imply

$$n_{j+vk} \leq n_j r^v + m_k \left( \sum_{l=0}^{v-1} r^l \right)$$

for  $v \geq 0$ . Thus

$$\begin{aligned} n_{j+vk} &\leq n_j r^v + m_k / (1 - r) \\ &\leq j m_1 r^v + m_k / (1 - r) \end{aligned}$$

by (6.4). But any positive integer  $l$  can be represented as  $l = vk + j$  for some  $v \geq 0$  and  $0 \leq j < k$ . Thus

$$n_l \leq (k-1)m_1 + m_k / (1 - r) = D$$

for all  $l \geq 1$ .

Q.E.D.

#### b. Proof of Corollary 2.5

Under the hypotheses of the corollary,

Norman

57.

$$P_s^\infty(A^L) = \sum_{i=1}^N P_{s_i}^{(1)}(A^L) \gamma_i(s) = 0.$$

Thus (2.12) implies

$$\|P_s^{(n)}(A^L)\| \leq L\alpha^n,$$

for  $n \geq 1$ . Thus the series  $\sum_{n=1}^{\infty} P_s^{(n)}(A^L)$  converges in the norm  $\|\cdot\|$  to an element of CL, and

$$\left\| \sum_{n=1}^{\infty} P_s^{(n)}(A^L) \right\| \leq \sum_{n=1}^{\infty} \|P_s^{(n)}(A^L)\| \leq L\alpha/(1-\alpha). \quad (6.6)$$

Let

$$X_n = \begin{cases} 1 & \text{if } (E_n, \dots, E_{n+L-1}) \in A^L \\ 0 & \text{if } (E_n, \dots, E_{n+L-1}) \notin A^L. \end{cases}$$

Then  $X = \sum_{n=1}^{\infty} X_n$  and  $E_s[X_n] = P_s^{(n)}(A^L)$ , so

$$E_s[X] = \sum_{n=1}^{\infty} P_s^{(n)}(A^L), \quad (6.7)$$

for all  $s \in S$ . This, in combination with (6.6), gives (2.14).

Clearly

$$X(s) = E_s[X_1] + E_s\left[\sum_{n=2}^{\infty} X_n\right]$$

$$= P_s^{(1)}(A^L) + E_s\left[E\left[\sum_{n=2}^{\infty} X_n \mid S_2\right]\right]$$

$$= P_s^{(1)}(A^1) + E_s[\chi(s_2)],$$

and  $\chi(s_i) = 0$ ,  $i = 1, \dots, N$ . If  $\chi'$  is another continuous function satisfying these conditions then  $\Delta = \chi - \chi' \in C(S)$  with  $U\Delta = \Delta$  and  $\Delta(s_i) = 0$ ,  $i = 1, \dots, N$ . Thus  $\Delta(s) \equiv 0$  by Theorem 2.3. Q.E.D.

c. Proof of Theorem 2.6

As was remarked in Section 2, a distance diminishing model can be regarded as an example of a homogeneous random system with complete connections. In the notational style of this paper (Iosifescu's is somewhat different), a homogeneous random system with complete connections is a system  $((S, \mathcal{B}), (E, \mathcal{A}), f, \tilde{\varphi})$  such that  $(S, \mathcal{B})$  and  $(E, \mathcal{A})$  are measurable spaces,  $f(\cdot)$  is a measurable mapping of  $E \times S$  into  $S$ ,  $\tilde{\varphi}_s(\cdot)$  is a probability measure on  $\mathcal{A}$  for each  $s \in S$ , and  $\tilde{\varphi}_A(\cdot)$  is a measurable real valued function on  $S$  for each  $A \in \mathcal{A}$ . An associated stochastic process  $\{E_n\}$  for such a system satisfies

$$H3' \quad P_s(E_1 \in A) = \tilde{\varphi}_A(s) \text{ and}$$

$$P_s(E_{n+1} \in A | E_j = e_j, 1 \leq j \leq n) = \tilde{\varphi}_A(f_{e_1 \dots e_n}(s))$$

for  $n \geq 1$ ,  $s \in S$ , and  $A \in \mathcal{A}$ . Under H4,  $\mathcal{B}$  can be taken to be the Borel subsets of  $S$ . Under H2,  $\mathcal{A}$  can be taken to be all subsets of  $E$ , and  $\tilde{\varphi}_s(\cdot)$  is a probability measure on  $\mathcal{A}$  if and only if

there is a non-negative real valued function  $\varphi_e(s)$  on  $E$  such that

$$\tilde{\varphi}_A(s) = \sum_{e \in A} \varphi_e(s)$$

and  $\sum_{e \in E} \varphi_e(s) = 1$ . Then H3' is equivalent to H3. Also, the measurability requirements on  $f$  and  $\tilde{\varphi}_A$  are weaker than continuity of  $f_e(\cdot)$  and  $\varphi_e(\cdot)$ , which are, in turn, weaker than H6 and H7.

The following lemma was proved (but not stated formally) by Iosifescu (1963, Chapter 3, Section 3).

Lemma 6.1. If a homogeneous random system with complete connections has the property that there is a sequence  $\{\varepsilon_n\}$  of positive numbers with  $\sum_{n=1}^{\infty} n\varepsilon_n < \infty$  and, for every  $l \geq 1$ , a probability measure  $P^{\infty}(A^l)$  on  $\mathcal{A}^l$  such that

$$|P_s^{(n)}(A^l) - P^{\infty}(A^l)| < \varepsilon_n$$

for all  $s \in S$ ,  $n$ ,  $l \geq 1$ , and  $A^l \in \mathcal{A}^l$ , then all of the conclusions of Theorem 2.6 apply to the associated stochastic process  $\{E_n\}$ .

The quantity  $P_s^{(n)}(A^l)$  is defined in (2.11).

For an ergodic model the probability measure  $P_s^{\infty}(A^l) = P^{\infty}(A^l)$  defined in (2.13) does not depend on  $s$ . Therefore (2.12) implies the hypotheses of Lemma 6.1 with  $\varepsilon_n = L\alpha^n$ , and the conclusions of Theorem 2.6 follow.

## APPENDIX: MODELS WITH FINITE STATE SPACES

a. Theory

The following definition is analogous to Definition 1.1.

Definition. A system  $(S, E, f, \phi)$  is a finite state model if  $S$  and  $E$  are finite sets,  $f, (\cdot)$  is a mapping of  $E \times S$  into  $S$ , and  $\phi, (\cdot)$  is a mapping of  $E \times S$  into the non-negative real numbers such that  $\sum_{e \in E} \phi_e(s) = 1$ .

Definition 1.2 defines associated stochastic processes  $\{S_n\}$  and  $\{E_n\}$  for any such model. It is possible to develop, by the methods of Sections 5 and 6, a theory of finite state models that completely parallels the theory of distance diminishing models surveyed in Section 2. This will not be done here, since the results concerning states obtained by these relatively complicated methods are, if anything, slightly inferior to those that can be obtained by applying the well known theory of finite Markov chains (see Kemeny and Snell (1960) and Feller (1957)) to the process  $\{S_n\}$ . However, the results concerning events in the ergodic case are new and important. Therefore, a development will be presented that leads to the latter results as directly as possible. Applications to stimulus-sampling theory will be given in Subsection b.

The natural analogue of H9 for finite state models is

H9' For any  $s, s' \in S$ ,  $T_n(s) \cap T_n(s')$  is not empty if  $n$  is sufficiently large.

This is equivalent to

H9' The finite Markov chain  $\{S_n\}$  has a single ergodic set, and this set is regular.

The terminology for finite Markov chains used in this appendix follows Kemeny and Snell (1960). By analogy with Definition 2.2, a finite state model that satisfies H9' will be called ergodic. The reader should note, however, that the associated process  $\{S_n\}$  need not be an ergodic Markov chain, since it may have transient states. If there happen not to be any transient states, the chain is ergodic and regular.

Lemma 1 is analogous to Theorem 2.3.

Lemma 1. For any ergodic finite state model there are constants  $C < \infty$  and  $\alpha < 1$  and a probability distribution  $K^\infty$  on  $S$ , such that

$$|E[\psi(S_n)] - E[\psi(S_\infty)]| \leq C\alpha^n |\psi| \quad (1)$$

for all real valued functions  $\psi$  on  $S$  and  $n \geq 1$ , where

$$E[\psi(S_\infty)] = \sum_{s \in S} \psi(s) K^\infty(\{s\}). \quad (2)$$

Proof. Let  $N$  be the number of states. To facilitate the use of matrix notation the states are denoted  $s_1, s_2, \dots, s_N$ .

The transition matrix  $P$  and the column vector  $\psi^*$  corresponding to  $\psi$  are then defined by

$$P_{ij} = K(s_i, \{s_j\}) \text{ and } \psi_i^* = \psi(s_i). \quad (3)$$

Then

$$E_{s_i} [\psi(S_n)] = (P^{n-1} \psi^*)_i \quad (4)$$

for  $n \geq 1$ .

There is a stochastic matrix  $A$ , all of whose rows are the same, say  $(a_1, \dots, a_N)$ , and there are constants  $b < \infty$  and  $\alpha < 1$  such that

$$|(P^{n-1})_{ij} - A_{ij}| \leq b\alpha^{n-1} \quad (5)$$

for all  $n \geq 1$  and  $1 \leq i, j \leq N$ . When  $\{S_n\}$  is regular, this assertion is Corollary 4.1.5 of Kemeny and Snell (1960). When  $\{S_n\}$  has transient states, that Corollary can be supplemented by Kemeny and Snell's Corollary 3.1.2 and a straightforward additional argument to obtain (5). Let  $K^\infty$  be the probability measure on  $S$  with  $K^\infty(\{s_j\}) = a_j$ , and let  $E[\psi(S_\infty)]$  be any coordinate of  $A\psi^*$ . Then (2) holds and

$$|E_{s_i} [\psi(S_n)] - E[\psi(S_\infty)]| = |(P^{n-1} \psi^*)_i - (A\psi^*)_i|$$

$$\begin{aligned}
 &= \left| \sum_{j=1}^N [(P^{n-1})_{ij} - A_{ij}] \psi_j^* \right| \\
 &\leq \sum_{j=1}^N |(P^{n-1})_{ij} - A_{ij}| |\psi(s_j)| \\
 &\leq Nb\alpha^{n-1} |\psi|.
 \end{aligned}$$

This gives (1) with  $C = Nb/\alpha$ .

Q.E.D.

The next lemma parallels Theorem 2.4.

Lemma 2. For any ergodic finite state model

$$|P^{(n)}(A^l) - P^\infty(A^l)| \leq C\alpha^n \quad (6)$$

for all  $n, l \geq 1$  and  $A^l \subset E^l$ , where

$$P^l(A^l) = \sum_{s \in S} P_s^{(1)}(A^l) K^\infty(s), \quad (7)$$

and  $C$  and  $\alpha$  are as in Lemma 1.

Proof. Just as in the proof of Theorem 2.4, (6.1) and (6.2) hold. Thus (6) follows from Lemma 1. Q.E.D.

Theorem 1 is the main result of this subsection.

Theorem 1. All of the conclusions of Theorem 2.6 hold for any ergodic finite state model.

Proof. A finite state model can be regarded as a homogeneous random system with complete connections, in the same sense that a distance diminishing model can be so regarded (see the first paragraph of Subsection 6c -- if  $\mathcal{S}$  and  $\mathcal{E}$  are taken to be the collections of all subsets of  $S$  and  $E$ , respectively, the measurability conditions in the definition of such a system evaporate). Thus Lemma 6.1 is applicable, and Theorem 1 follows from Lemma 2. Q.E.D.

b. Application to Stimulus-Sampling Theory

Consider the general two-choice situation described in the first paragraph of Subsection 3a. The state  $S_n$  at the beginning of trial  $n$  for the  $N$  element component model with fixed sample size  $v$  (Estes (1959)) can be taken to be the number of stimulus elements conditioned to response  $A_1$  at the beginning of the trial. Thus

$$S = \{0, 1, \dots, N\}, \quad (8)$$

a finite set. The event space  $E$  can be taken to be

$$E = \{(i, j, k, l): 0 \leq i \leq v, 1 \leq j, k \leq 2, 0 \leq l \leq 1\} \quad (9)$$

where  $i$  is the number of stimulus elements in the trial sample conditioned to  $A_2$ ,  $A_j$  is the response and  $O_{jk}$  the trial outcome, and  $l = 1$  or  $0$  depending on whether or not conditioning is effective. The corresponding event operators can be written

$$f_{ij11}(s) = s + \min(i, N - s), \quad (10)$$

$$f_{ij21}(s) = s - \min(v - i, s), \text{ and} \quad (11)$$

$$f_{ijk0}(s) = s. \quad (12)$$

Of course,  $i$  elements conditioned to  $A_2$  cannot be drawn if  $i > N - s$ , so the definition of  $f_{ij11}(s)$  is irrelevant in this case. The definition given in (10) makes  $f_{ij11}(\cdot)$  monotonic throughout  $S$ . The same holds for (11). The operators given by (10) and (11) are, incidentally, analogous in form to the linear operators

$$f(p) = p + \theta(1 - p) \text{ and } g(p) = p - \theta p,$$

if taking the minimum of two numbers is regarded as analogous to multiplying them. Finally, the corresponding operator application probabilities are

$$\varphi_{ijk\ell}(s) = \frac{\binom{N-s}{i} \binom{s}{v-i}}{\binom{N}{v}} \left( \frac{v-i}{v} \right)_j \pi_{jk} [c_{jk}]_{\ell} \quad (13)$$

where  $(p)_i = \delta_{i1}p + \delta_{i2}(1-p)$ ,  $[p]_{\ell} = \delta_{\ell1}p + \delta_{\ell0}(1-p)$ ,  $c_{jk}$  is to be interpreted as the probability that conditioning is effective if outcome  $O_{jk}$  occurs, and  $\left( \frac{J}{m} \right)$  is 0 unless  $0 \leq m \leq J$ .

For any choice of  $N \geq v \geq 1$  and  $0 \leq \pi_{ij}, c_{ij} \leq 1$ , (8) - (13) define a finite state model that will be referred to below as a

fixed sample size model.

**Theorem 2.** A fixed sample size model with  $\pi_{ij}$  and  $c_{ij}$  positive for all  $1 \leq i, j \leq 2$  is ergodic.

**Proof.** It is clear that if  $S_n < N$  then the sample on trial  $n$  will contain elements conditioned to  $A_2$  with positive probability. Given such a sample,  $A_{2,n}$  will occur and be followed by  $O_{21,n}$  with positive probability, and conditioning will be effective with positive probability. Thus  $S_{n+1} > S_n$  with positive probability, and it follows that the state  $N$  can be reached from any state. Thus there is only one ergodic set, and it contains  $N$ . Furthermore, if  $S_n = N$  then  $A_{1,n}$  occurs with probability 1, and  $O_{11,n}$  follows with positive probability. So  $S_{n+1} = N$  with positive probability, and the ergodic set is regular. Q.E.D.

It follows from Theorems 1 and 2 that the conclusions of Theorem 2.6 are available for any fixed sample size model with  $0 < \pi_{ij}, c_{ij}$  for all  $i$  and  $j$ . Letting  $D$  be the subset

$$D = \{(i, j, k, \ell) : j = 1\}$$

of  $E$ , and  $h_n = h(E_n)$ , where  $h$  is the indicator function of  $D$ , the conclusions of Theorem 2.6 include a law of large numbers and, possibly, a central limit theorem for the number  $S_{m,n} = \sum_{j=m}^{m+n-1} A_j$  of  $A_1$  responses in the  $n$  trial block starting on trial  $m$ . A simple expression for  $\sigma^2 = \sigma_h^2$  can be readily calculated for the pattern model ( $v = 1$ ) with equal  $c_{ij}$  under noncontingent reinforcement.

**Theorem 3.** A fixed sample size model with  $v = 1$ ,  $0 < \pi_{11} = \pi_{21} = \pi_1 < 1$ , and  $c_{ij} = c > 0$  is ergodic. The law of large numbers

$$\lim_{n \rightarrow \infty} E_s [((1/n)S_{m,n} - \pi_1)^2] = 0$$

and central limit theorem

$$\lim_{n \rightarrow \infty} P\left(\frac{S_{m,n} - n\pi_1}{\sqrt{n} \sigma} < x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp(-t^2/2) dt$$

hold, where

$$\sigma^2 = \pi_1 (1 - \pi_1) (1 + 2(1 - c)/c).$$

Proof. That

$$E[h(E_\infty)] = \lim_{n \rightarrow \infty} P(A_{1,n}) = \pi_1$$

follows from (37) in Atkinson and Estes (1963). The value of  $\sigma^2$  can be obtained from (2.17) and Atkinson and Estes' formula (41). E.D.

The methods of this subsection are equally applicable to the component model with fixed sampling probabilities (Estes (1959)).

## REFERENCES

- Atkinson, R. C., and Estes, W. K. Stimulus sampling theory. In R. D. Luce, R. R. Bush, and E. Galanter (Eds.), Handbook of mathematical psychology, vol. II. New York: Wiley, 1963. Pp. 121-268.
- Breiman, L. A strong law of large numbers for a class of Markov chains. Annals of Mathematical Statistics, 1960, 31, 801-803.
- Bush, R. R. Identification learning. In R. D. Luce, R. R. Bush, and E. Galanter (Eds.), Handbook of mathematical psychology, vol. III. New York: Wiley, 1965. Pp. 161-203.
- Bush, R. R. and Mosteller, F. Stochastic models for learning. New York: Wiley, 1955.
- Doebelin, W. and Fortet, R. Sur des chaînes à liaisons complètes. Bulletin de la Société Mathématique de France, 1937, 65, 132-148.
- Estes, W. K. Component and pattern models with Markovian interpretations. In R. R. Bush and W. K. Estes (Eds.), Studies in mathematical learning theory. Stanford: Stanford University Press, 1959. Pp. 9-52.
- Estes, W. K., and Suppes, P. Foundations of linear models. In R. R. Bush and W. K. Estes (Eds.), Studies in mathematical learning theory. Stanford: Stanford University Press, 1959. Pp. 137-179.
- Feller, W. An introduction to probability theory and its applications, vol. I, 2nd ed. New York: Wiley, 1957.

- Peller, W. An introduction to probability theory and its applications, vol. II. New York: Wiley, 1966.
- Ionescu Tulcea, C. On a class of operators occurring in the theory of chains of infinite order. Canadian Journal of Mathematics, 1959, 11, 112-121.
- Ionescu Tulcea, C., and Marinescu, G. Theorie ergodique pour des classes d'operations non completement continues. Annals of Mathematics, 1950, 52, 140-147.
- Iosifescu, M. Random systems with complete connections with an arbitrary set of states. Revue de Mathématique Pures et Appliquées, 1963, 8, 611-645. A translation of this paper from the Russian can be obtained for \$7.50 from Addis Translations, 129 Pope Street, Menlo Park, California 94025.
- Iosifescu, M. and Theodorescu, R. On Bush-Mosteller stochastic models for learning. Journal of Mathematical Psychology, 1965, 2, 196-203.
- Jamison, B. Asymptotic behavior of successive iterates of continuous functions under a Markov operator. Journal of Mathematical Analysis and Applications, 1964, 9, 203-214.
- Jamison, B. Ergodic decompositions induced by certain Markov operators. Transactions of the American Mathematical Society, 1965, 117, 451-466.
- Karlin, S. Some random walks arising in learning models I. Pacific Journal of Mathematics, 1953, 3, 725-756.

Kemeny, J. G., and Snell, J. L. Finite Markov chains. Princeton: Van Nostrand, 1960.

Kennedy, P. A convergence theorem for a certain class of Markov processes. Pacific Journal of Mathematics, 1957, 7, 1107-1124.

Lamperti, J. and Suppes, P. Chains of infinite order and their application to learning theory. Pacific Journal of Mathematics, 1959, 9, 739-754.

Lamperti, J. and Suppes, P. A correction to "Chains of infinite order and their application to learning theory." Pacific Journal of Mathematics, 1965, 15, 1471-1472.

Lovejoy, E. Analysis of the overlearning reversal effect. Psychological Review, 1966, 73, 87-103.

Norman, M. F. Incremental learning on random trials. Journal of Mathematical Psychology, 1964, 1, 336-350.

Onicescu, O. and Mihoc, G. Sur les chaînes de variables statistiques. Bulletin de la Société Mathématique de France, 1935, 59, 174-192.

Rose, R. M. Models for experiments with two complimentary reinforcing events. Unpublished doctoral dissertation, University of Pennsylvania, 1964.

Wyckoff, L. B. The role of observing responses in discrimination learning, part I. Psychological Review, 1952, 59, 431-442.

Yellott, J. I. Some effects of noncontingent success in human probability learning. Technical Report No. 89, 1965. Institute for Mathematical Studies in the Social Sciences, Stanford University.

Norman

81.

FOOTNOTE

<sup>1</sup>This research was supported in part by National Science Foundation grant NSF GU 1923-3.

ON THE LINEAR MODEL

WITH TWO ABSORBING BARRIERS

M. Frank Norman

University of Pennsylvania, Philadelphia, Pennsylvania

To be published in the Journal of Mathematical Psychology

### Abstract

A family of linear models for learning in two choice situations is considered. These models have in common the assumption that nonreward has no effect on response probability. The function  $\gamma(p)$  that relates asymptotic probability of one of the responses to its initial probability is studied intensively. It is shown to be closely related to the total number  $\chi(p)$  of response alternations. The asymptotic probability of making the less favorable response is shown to be small when the learning rates associated with reward are small. Finally, some of the basic analytic function theoretic properties of  $\gamma(p)$  are presented.

### 1. Introduction

Throughout this section we consider a two choice ( $A_1$  or  $A_2$ ) animal learning experiment where, on any trial, response  $A_1$  is followed by reward (event  $E_1$ ) with probability  $\pi_1 > 0$ . Occurrence of  $A_i$  or  $E_i$  on trial  $n$  is denoted  $A_{i,n}$  or  $E_{i,n}$ , and a subject's probability of  $A_{1,n}$  is denoted  $p_n$ . When correction or retracing is permitted after a nonreinforced response (an "error"), asymptotic  $A_1$  response probability is found to depend on a number of factors, among them the species of the subject and the type of discrimination required (Bitterman, 1965). In such experiments reward has always been set up for exactly one of the two responses before each trial so that reinforcement is noncontingent, i.e.  $\pi_1 = 1 - \pi_2 = \pi$ . Noncorrection experiments using both contingent and noncontingent reinforcement, on the other hand, have consistently yielded very high asymptotic proportions of choices of the response with the highest probability of being rewarded (Bitterman, Woodinsky, and Candland, 1958; Behrend and Bitterman, 1961; Brody, 1965; Mayer, 1960; Parducci and Polt, 1958; Stanley, 1950; and Weinstock, North, Brody, and LoGuidice, 1965).

The possibility has often been considered that the effect of nonreinforcement in these experiments might be nil or practically nil. This assumption, embedded within linear learning models, has interesting consequences. When coupled with the auxiliary

assumption that a reinforced correction response has the same effect on response probability as a reinforced original response, it leads to the transition rules

$$p_{n+1} = \begin{cases} (1-\theta)p_n + \theta & \text{if } A_{1,n}E_{1,n} \text{ or } A_{2,n}\tilde{E}_{2,n} \\ (1-\theta)p_n & \text{if } A_{2,n}E_{2,n} \text{ or } A_{1,n}\tilde{E}_{1,n} \end{cases} \quad (1)$$

for  $p_n$ . The probabilities that the first and second rows are applicable are easily seen to be  $\pi$  and  $1-\pi$  respectively. This model is quite useful psychologically and quite well understood mathematically. It predicts probability matching:

$$\lim_{n \rightarrow \infty} p_p(A_{1,n}) = \pi$$

for any value  $p$  of  $p_1$ .

The corresponding linear model for the noncorrection experiment has transition equations

$$p_{n+1} = \begin{cases} (1-\theta_1)p_n + \theta_1 & \text{if } A_{1,n}E_{1,n} \\ (1-\theta_2)p_n & \text{if } A_{2,n}E_{2,n} \\ p_n & \text{if } A_{1,n}\tilde{E}_{1,n} \text{ or } A_{2,n}\tilde{E}_{2,n} \end{cases} \quad (2)$$

$1 \geq \theta_1, \theta_2 > 0$ . For the sake of the present discussion we suppose that  $\theta_1 = \theta_2 = \theta$ , but the more general model, which might arise, for instance, if the magnitudes or delays of the rewards  $E_1$  and  $E_2$

were unequal, is also treated below. The identity operator or two absorbing barrier linear model (2) has been considered by many investigators (Bush and Mosteller, 1955, pp. 291-294; Brody, 1965;

Weinstock, et al, 1965) in this setting with considerable success. (See the last two paragraphs of this paper for an observation that suggests that the model may have been more successful than the authors of the latter paper realized.) There are no special difficulties in deriving predictions from this model either by mathematical approximation techniques (Bush and Mosteller, 1955, pp. 286-291; Mosteller and Tatsuoka, 1960) or by Monte Carlo methods (Brody, 1965; Weinstock, et al, 1965). Nevertheless there are some large gaps in our knowledge about this model. The purpose of this paper is to fill some of these.

The quantities

$$\gamma(p) = \lim_{n \rightarrow \infty} P_p(A_{1,n}) \quad (3)$$

and

$$\chi(p) = E_p[\text{total number of response alternations}] \quad (4)$$

are of basic interest. There are practically no cases (i.e. special parameter values) for which simple formulas for  $\gamma$  and  $\chi$  are known. The research reported below began as an attempt to provide the first proof that, in the equal theta case, the asymptotic probability of choosing the unfavorable side is small, at least when  $\theta$  is small; that is,  $\gamma(p) \rightarrow 0$  as  $\theta \rightarrow 0$  if  $\pi_2 > \pi_1$ ,  $\theta_1 = \theta_2 = \theta$ , and  $0 \leq p < 1$ . It developed into a fairly general investigation of the functions  $\gamma$  and  $\chi$ . The lemmas of Section 2

provide the foundation for the subsequent development. The main result of Section 3 is a relation between  $\chi$  and  $\gamma$  that essentially reduces the study of the former to that of the latter. The two special cases  $\pi_1 = \pi_2 = 1$  and  $\theta_1 = \theta_2$  of this result are easily derivable from formulas relating the asymptotic probability of  $A_1$  to the total number of runs of  $A_1$ 's that Bush (1959, Sections 5 and 6) obtained by another method. In Section 4 bounds are obtained for  $\gamma$  that are sufficiently precise to permit deduction that the asymptotic probability of choosing the unfavorable side is small when the learning rates are small. The theorem of Section 5 is concerned with the analytic character of  $\gamma$ . Most of the results of that section are extensions or refinements of those of Karlin (1953, Sections 1 and 5), who concentrated on the case  $\pi_1 = \pi_2 = 1$ .

## 2. Fundamentals

Throughout the remainder of the paper we will be concerned with the identity operator model (2) under the conditions

$$1 \geq \theta_1, \theta_2, \pi_1, \pi_2 > 0.$$

For any (real valued) function  $\psi$  on  $[0,1]$  we define  $|\psi|$  by

$$|\psi| = \sup_{0 \leq p \leq 1} |\psi(p)|.$$

We let  $D$  be the class of all differentiable functions with bounded derivative on  $[0,1]$ . Such functions are necessarily continuous. If  $\psi \in D$ , its norm  $\|\psi\|$  is defined by

$$\|\psi\| = |\psi| + |\psi'|.$$

7.

Thus the relation  $\lim_{n \rightarrow \infty} \|\psi_n - \psi\| = 0$  for  $\psi_n, \psi \in D$  is tantamount to the uniform convergence of  $\psi_n$  to  $\psi$  and of  $\psi'_n$  to  $\psi'$ . The operator

$$U\psi(p) = E[\psi(p_{n+1}) | p_n = p], \text{ i.e.}$$

$$U\psi(p) = \psi((1-\theta_1)p + \theta_1)\pi_1 p + \psi((1-\theta_2)p)\pi_2(1-p) + \psi(p)(1-\pi_1 p - \pi_2(1-p)) \quad (5)$$

maps  $D$  into  $D$  and is linear and positive (that is, it preserves nonnegative functions). It can be shown by a simple inductive argument that

$$U^n \psi(p) = E_p[\psi(p_{n+1})], \quad (6)$$

$0 \leq p \leq 1$ , where  $U^n$  is the  $n^{\text{th}}$  iterate of the operator  $U$ .

Lemmas 1 and 2 below follow easily from the results of Norman (1967, see especially Subsection a of Section 3).

Lemma 1. The only absorbing states (barriers) of the Markov process  $(p_n)_{n=1}^{\infty}$  are 0 and 1. The process converges with probability 1 to a random absorbing state  $p_{\infty}$ . The sequence  $(P_p(A_{1,n}))_{n=1}^{\infty}$  converges as  $n \rightarrow \infty$  to

$$\gamma(p) = P_p(p_{\infty} = 1). \quad (7)$$

The function  $\gamma$  belongs to  $D$  and is the only continuous solution of the functional equation

$$U\gamma = \gamma$$

(8)

that has the boundary values  $\gamma(0) = 0$  and  $\gamma(1) = 1$ . The function

$$\gamma_{a,b}(p) = a(1-\gamma(p)) + b\gamma(p)$$

is the only continuous function such that  $U\gamma_{a,b} = \gamma_{a,b}$ ,

$\gamma_{a,b}(0) = a$ , and  $\gamma_{a,b}(1) = b$ . There are  $\alpha < 1$  and  $C < \infty$  such that

$$\|E[\psi(p_n)] - E[\psi(p_{n-1})]\| \leq C\alpha^n \|\psi\| \quad (9)$$

for all  $\psi \in D$  and  $n \geq 1$ .

**Lemma 2.** The total number  $Y$  of alternations between responses is finite with probability 1. In fact, for any  $0 \leq p \leq 1$ ,  $E_p[Y] < \infty$ . The function  $\chi(p) = E_p[Y]$  belongs to  $D$  and is the unique continuous solution of the functional equation

$$\chi(p) = (2-\theta_1\pi_1-\theta_2\pi_2)p(1-p) + U\chi(p) \quad (10)$$

for which  $\chi(0) = \chi(1) = 0$ .

Let  $X_n$  be the indicator random variable for the event  $A_{1,n}$ .

Since

$$Y = \sum_{n=1}^{\infty} |X_{n+1} - X_n| < \infty$$

it follows that  $\{X_n\}$  converges with probability 1. Let  $X_\infty$  be the limiting random variable. Clearly  $X_\infty \in (0,1)$  with probability 1,

and it is plausible that  $X_\infty = p_\infty$  with probability 1. To prove this we note that

$$\begin{aligned}
 P_p(X_\infty \neq p_\infty) &= P_p(X_\infty = 1, p_\infty = 0) + P_p(X_\infty = 0, p_\infty = 1) \\
 &= E_p[X_\infty(1-p_\infty)] + E_p[(1-X_\infty)p_\infty] \\
 &= \lim_{n \rightarrow \infty} E_p[X_n(1-p_n)] + \lim_{n \rightarrow \infty} E_p[(1-X_n)p_n] \\
 &= \lim_{n \rightarrow \infty} E_p[E[X_n(1-p_n)|p_n]] + \lim_{n \rightarrow \infty} E_p[E[(1-X_n)p_n|p_n]] \\
 &= \lim_{n \rightarrow \infty} E_p[(1-p_n)E[X_n|p_n]] + \lim_{n \rightarrow \infty} E_p[p_n(1-E[X_n|p_n])] \\
 &= 2 \lim_{n \rightarrow \infty} E_p[(1-p_n)p_n] = 2E_p[(1-p_\infty)p_\infty] = 0.
 \end{aligned}$$

Therefore  $\gamma(p)$  has the following behavioral interpretation:

$$\gamma(p) = P_p(X_\infty = 1).$$

Since  $X_n$  is an indicator,  $X_\infty = 1$  means that  $X_n = 1$  for all but a finite number of  $n$ , i.e.  $A_{1,n}$  for all but a finite number of  $n$ .

A standard notation for the latter event is  $\liminf A_{1,n}$ . Therefore  $\gamma(p) = P_p(\liminf A_{1,n})$ .

If  $\psi \in D$  with  $\psi(0) = 0$  and  $\psi(1) = 1$  (e.g.  $\psi(p) = p$ ) then

$$E_p[\psi(p_\infty)] = \psi(0)(1-\gamma(p)) + \psi(1)\gamma(p) = \gamma(p).$$

Thus (9) and (6) imply that  $\|U^n \psi - \gamma\|$  converges geometrically to 0 as  $n \rightarrow \infty$ . This gives us an iterative method of approximating  $\gamma$  that should be useful in numerical computations.

The functions  $\gamma$  and  $\chi$  and the operator  $U$  depend on four parameters. When it is necessary to call attention to the dependence on some of these parameters we use notations such as  $\gamma(p; \pi_1, \pi_2)$ ,  $\gamma(p; \theta_1, \theta_2, \pi_1, \pi_2)$ , and  $U_{\theta_1, \theta_2, \pi_1, \pi_2}$ . The study of the dependence of  $\gamma$  upon its parameters is considerably simplified by the following lemma which states the obvious fact that the probability of absorption on response  $A_1$  when the parameters are  $\theta_1, \theta_2, \pi_1$ , and  $\pi_2$  and  $p$  is the initial probability of  $A_1$  is the same as the probability of absorption on  $A_2$  when the parameters are  $\theta_2, \theta_1, \pi_2$ , and  $\pi_1$  and  $p$  is the initial probability of  $A_2$  so that  $1-p$  is the initial probability of  $A_1$ .

Lemma 3.

$$\gamma(p; \theta_1, \theta_2, \pi_1, \pi_2) = 1 - \gamma(1-p; \theta_2, \theta_1, \pi_2, \pi_1).$$

Another proof of this equality is obtained by verifying that the function

$$\phi(p) = 1 - \gamma(1-p; \theta_2, \theta_1, \pi_2, \pi_1)$$

belongs to  $D$ , and satisfies the functional equation

$$U_{\theta_1, \theta_2, \pi_1, \pi_2} \phi = \phi$$

and the boundary conditions  $\phi(0) = 0$ ,  $\phi(1) = 1$ . We use a similar method to prove the following lemma, which is only slightly less obvious than the preceding one.

Lemma 4. For any  $0 < x \leq 1/\max(\pi_1, \pi_2)$

$$\gamma(p; x\pi_1, x\pi_2) = \gamma(p; \pi_1, \pi_2).$$

In other words,  $\gamma$  depends on  $\pi_1$  and  $\pi_2$  only through their ratio.

Proof. The functional equation (8) for  $\gamma(p; \pi_1, \pi_2)$  is equivalent to

$$\begin{aligned} (\pi_1 p + \pi_2(1-p))\gamma(p; \pi_1, \pi_2) \\ = \pi_1 p \gamma((1-\theta_1)p + \theta_1; \pi_1, \pi_2) + \pi_2(1-p)\gamma((1-\theta_2)p; \pi_1, \pi_2). \end{aligned}$$

Multiplying both sides by  $x$  we see that  $\gamma(p; \pi_1, \pi_2)$  satisfies an equation equivalent to the functional equation for  $\gamma(p; x\pi_1, x\pi_2)$ . O.E.D.

In particular, the absorption probability  $\gamma(p; v, v)$  for the case  $\pi_1 = \pi_2 = v$  does not depend on  $v$ . \*

We now describe a method that permits us to obtain bounds on the solutions  $\gamma$  and  $\chi$  of the functional equations (8) and (10) by solving corresponding functional inequalities.

Definition. A function  $\psi$  on  $[0, 1]$  is superregular (regular, subregular) if and only if  $\psi(p) \geq$  ( $=$ ,  $\leq$ )  $U\psi(p)$  for all  $0 \leq p \leq 1$ .

These concepts are standard in the potential theory of Markov processes (see, for instance, Kemeny, Snell, and Knapp (1966)).

The next lemma shows the usefulness of these notions and justifies the terminology.

Lemma 5. Let  $\psi \in D$  be superregular (subregular) with  $\psi(0) = a$  and  $\psi(1) = b$ . Then  $\psi(p) \geq (\leq) \gamma_{a,b}(p)$ , where  $\gamma_{a,b}(p)$  is the continuous regular function with  $\gamma_{a,b}(0) = a$  and  $\gamma_{a,b}(1) = b$ .

Proof. Since  $\psi(p) \geq (\leq) U\psi(p)$  and  $U^n$  is a positive linear operator,  $U^n\psi(p) \geq (\leq) U^{n+1}\psi(p)$  for all  $n \geq 0$ . But

$$\lim_{n \rightarrow \infty} U^n\psi(p) = E_p[\psi(p_\infty)] = \gamma_{a,b}(p)$$

by Lemma 1. Thus  $\psi(p) \geq (\leq) \gamma_{a,b}(p)$  for all  $0 \leq p \leq 1$ . Q.E.D.

Lemma 5 implies the slightly more general Lemma 6.

Lemma 6. Suppose that  $\psi$ ,  $\phi$ , and  $g \in D$ , all three functions vanish at 0 and at 1, and

$$\psi(p) \geq (\leq) g(p) + U\psi(p),$$

while

$$\phi(p) = g(p) + U\phi(p).$$

Then  $\psi(p) \geq (\leq) \phi(p)$  for all  $0 \leq p \leq 1$ .

Proof. Let  $\Delta(p) = \psi(p) - \phi(p)$ . Then  $\Delta \in D$ ,  $\Delta$  is superregular (subregular), and  $\Delta(0) = \Delta(1) = 0$ . Hence by Lemma 5,

$$\Delta(p) \geq (\leq) \gamma_{0,0}(p) = 0$$

for all  $0 \leq p \leq 1$ .

Q.E.D.

We remark that Lemmas 5 and 6 generalize immediately to the general absorbing distance diminishing models treated by Norman (1967).

### 3. Alternations

Let  $I$  be the identity function on the real line:  $I(p) \equiv p$ , and let  $B(p) \equiv p(1-p)$ . A simple computation yields the following important result.

Lemma 7.  $UI = I + (\theta_1 \pi_1 - \theta_2 \pi_2)B$ .

Taking  $\theta_1 \pi_1 = \theta_2 \pi_2$  and using the characterization of  $\gamma$  given in Lemma 1, we obtain

Theorem 1. If  $\theta_1 \pi_1 = \theta_2 \pi_2$ , then  $\gamma = I$ .

When  $\theta_1 \pi_1 \neq \theta_2 \pi_2$  Lemma 7 leads to a relation between  $\gamma$  and  $\chi$ .

Theorem 2. If  $\theta_1 \pi_1 \neq \theta_2 \pi_2$ , then

$$\chi = \frac{2 - \theta_1 \pi_1 - \theta_2 \pi_2}{\theta_2 \pi_2 - \theta_1 \pi_1} (I - \gamma). \quad (11)$$

Proof. Let the function on the right be denoted  $F$ . Since  $I$  and  $\gamma$  are continuous,  $F$  is too. And since  $I$  and  $\gamma$  agree at 0 and 1,  $F(0) = F(1) = 0$ . Finally

$$\begin{aligned}
 UF &= \frac{2-\theta_1\pi_1-\theta_2\pi_2}{\theta_2\pi_2-\theta_1\pi_1} (UI - U\gamma) \\
 &= \frac{2-\theta_1\pi_1-\theta_2\pi_2}{\theta_2\pi_2-\theta_1\pi_1} (I + (\theta_1\pi_1-\theta_2\pi_2)B - \gamma)
 \end{aligned}$$

by (8) and Lemma 7. So

$$UF = F - (2-\theta_1\pi_1-\theta_2\pi_2)B,$$

which is equivalent to (10). Thus Lemma 2 implies  $\chi = F$ . Q.E.D.

By evaluating both sides of (11) at the random point  $p_1$  and taking expectations we obtain a relation

$$E[Y] = \frac{2-\theta_1\pi_1-\theta_2\pi_2}{\theta_2\pi_2-\theta_1\pi_1} (P(A_{1,1}) - P(p_\infty = 1))$$

between the three quantities  $E[Y]$ ,  $P(A_{1,1})$ , and  $P(p_\infty = 1)$  that does not depend on the distribution of  $p_1$ . This relation can be tested empirically once  $\theta_1$  and  $\theta_2$  have been estimated. Alternatively, when  $\theta_1 = \theta_2 = \theta$ , it can be used to estimate  $\theta$ . Since the absorption probabilities are not changed when  $\pi_1$  and  $\pi_2$  are both multiplied by the same constant, the quantity  $P(A_{1,1}) - P(p_\infty = 1)$  can be cancelled to obtain

$$\frac{E^{(x)}[Y]}{E^{(x')}[Y]} = \frac{x' (2 - x(\theta_1\pi_1 + \theta_2\pi_2))}{x (2 - x'(\theta_1\pi_1 + \theta_2\pi_2))},$$

where the reinforcement probabilities corresponding to  $E^{(x)}[Y]$  are  $z\pi_1$  and  $z\pi_2$ . In particular, if  $x$  and  $x'$  are small, or  $\theta_1$  and  $\theta_2$  are small, or some combination of the two,

$$\frac{E^{(x)}[Y]}{E^{(x')}[Y]} \approx \frac{x'}{x}.$$

In Section 4 we use (11) and our results on the asymptotic behavior of  $\gamma$  as the thetas become small to obtain an asymptotic expression for  $\chi$ .

The next lemma will help us derive bounds for  $\chi$  when  $\theta_1\pi_1 = \theta_2\pi_2$ .

Lemma 8. If  $\theta_1\pi_1 = \theta_2\pi_2$  then

$$(1 - \theta_1\theta_2\max(\pi_1, \pi_2))B(p) \leq UB(p) \leq (1 - \theta_1\theta_2\min(\pi_1, \pi_2))B(p).$$

Proof. A straightforward computation shows that, for any  $\theta_1, \theta_2, \pi_1$ , and  $\pi_2$

$$UB(p) = B(p)(1 - [\theta_1\pi_1 - \theta_2(1-\theta_2)\pi_2]p - [\theta_2\pi_2 - \theta_1(1-\theta_1)\pi_1](1-p)). \quad (12)$$

When  $\theta_1\pi_1 = \theta_2\pi_2$  this reduces to

$$UB(p) = B(p)(1 - \theta_2^2\pi_2p - \theta_1^2\pi_1(1-p)).$$

Since  $\theta_2^2\pi_2 = \theta_2\theta_1\pi_1$  and  $\theta_1^2\pi_1 = \theta_1\theta_2\pi_2$  the lemma follows. Q.E.D.

Theorem 3. If  $\theta_1\pi_1 = \theta_2\pi_2$ , then

$$\frac{2-\theta_1\pi_1-\theta_2\pi_2}{\theta_1\theta_2^{\max(\pi_1,\pi_2)}} B(p) \leq \chi(p) \leq \frac{2-\theta_1\pi_1-\theta_2\pi_2}{\theta_1\theta_2^{\min(\pi_1,\pi_2)}} B(p).$$

Proof. It follows from Lemma 8 that the function  $F(p)$  on the right (left) satisfies

$$F(p) \geq (\leq) (2-\theta_1\pi_1-\theta_2\pi_2)B(p) + UF(p).$$

Since these are the functional inequalities corresponding to (10), an application of Lemma 6 completes the proof. Q.E.D.

These bounds are tight when  $\pi_1 - \pi_2$  is small. In the important special case  $\pi_1 = \pi_2$  (and  $\theta_1 = \theta_2$ ) they yield  $\chi$  exactly.

Corollary. When  $\theta_1 = \theta_2 = \theta$  and  $\pi_1 = \pi_2 = v$ ,

$$\chi = \frac{2(1-\theta v)}{\theta^2 v} B.$$

#### 4. Small Learning Rates

In this section we will show that if  $\pi_1$  and  $\pi_2$  are fixed and  $(\theta_1, \theta_2)$  approaches  $(0,0)$  along a line in the  $\theta_1, \theta_2$  plane for which  $\theta_2\pi_2 > \theta_1\pi_1$ , then  $\gamma(p; \theta_1, \theta_2) \rightarrow 0$  for all  $0 \leq p < 1$ . Moreover the convergence is extremely rapid. The inequality  $\theta_2\pi_2 > \theta_1\pi_1$  may be thought of as indicating that  $\lambda_2$  is the most favorable response. When  $\theta_1 = \theta_2$  this inequality reduces to  $\pi_2 > \pi_1$

as, intuitively, it should. Analogous results for  $\theta_1 \pi_1 > \theta_2 \pi_2$  may be obtained by applying the results established below to  $\gamma(1-p; \theta_2, \theta_1, \pi_2, \pi_1)$  and using Lemma 3.

We assume throughout this section that  $(\theta_1, \theta_2)$  is confined to a line through the origin in the  $\theta_1, \theta_2$  plane on which  $\theta_2 \pi_2 > \theta_1 \pi_1$ . This line is characterized by the ratio  $\zeta = \theta_1 / \theta_2$ , and  $\theta_1$  and  $\theta_2$  are of the form

$$\theta_1 = \zeta \theta, \quad \theta_2 = \theta \quad (13)$$

$0 < \theta \leq \min(1, 1/\zeta)$ . Clearly  $\zeta > 0$  and, if

$$\omega = \pi_1 / \pi_2,$$

then

$$1 > \omega \zeta. \quad (14)$$

For any  $x, \theta > 0$ , let  $\psi_{x, \theta}$  be the function on  $[0, 1]$  defined by

$$\psi_{x, \theta}(p) = e^{xp/\theta}.$$

Most of our effort in this section goes into the proof of the following lemma.

Lemma 9. There are positive constants  $y = y(\omega, \zeta)$  and  $z = z(\omega, \zeta)$  such that  $\psi_{y, \theta}$  is subregular and  $\psi_{z, \theta}$  is superregular for all  $0 < \theta \leq \min(1, 1/\zeta)$ .

Proof. For any  $x > 0$  and  $0 \leq p \leq 1$

$$(u_{\theta_1, \theta_2} \psi_{x, \theta})(p)$$

$$= e^{xp/\theta} \left[ e^{x\theta_1(1-p)/\theta} \pi_1 p + e^{-x\theta_2 p/\theta} \pi_2(1-p) + (1 - \pi_1 p - \pi_2(1-p)) \right]$$

$$= \psi_{x, \theta}(p) \left[ 1 + (e^{x\zeta(1-p)} - 1) \pi_1 p - (1 - e^{-xp}) \pi_2(1-p) \right]$$

by (13). Thus  $\psi_{x, \theta}$  is subregular (superregular) for all  $0 < \theta \leq \min(1, 1/\zeta)$  if and only if

$$(e^{x\zeta(1-p)} - 1) \pi_1 p \geq (\leq) (1 - e^{-xp}) \pi_2(1-p) \quad (15)$$

for all  $0 \leq p \leq 1$ . (The reader should note that these inequalities do not involve  $\theta$ .) However the difference between the two sides of (15) is continuous throughout  $[0, 1]$ , therefore these inequalities hold throughout  $[0, 1]$  if and only if they hold throughout  $(0, 1)$ , i.e.

$$\frac{(e^{x\zeta(1-p)} - 1)}{(1 - e^{-xp})} \frac{p}{1-p} \geq (\leq) \frac{1}{\omega}$$

for all  $0 < p < 1$ . In terms of the function  $V$  on  $(-\infty, \infty)$  defined by

$$V(u) = \begin{cases} (e^u - 1)/u & \text{if } u \neq 0 \\ 1 & \text{if } u = 0, \end{cases} \quad (16)$$

$\psi_{x,\theta}$  is subregular (superregular) for all  $0 < \theta \leq \min(1, 1/\zeta)$  if and only if

$$f_{\zeta}(p, x) = \frac{V(x\zeta(1-p))}{V(-xp)} \geq (\leq) \frac{1}{\zeta\omega} \quad (17)$$

for all  $0 < p < 1$ .

Now

$$V(u) = \int_0^1 e^{uw} dw$$

for all real  $u$ . Since the integrand is positive,  $V(u) > 0$  for all real  $u$ . By taking the derivatives under the integral sign we see that

$$V^{(k)}(u) = \int_0^1 w^k e^{uw} dw$$

for  $k = 1$  or  $2$ . Thus  $V'(u) > 0$  for all real  $u$ , so  $V$  is strictly increasing. It follows that

$$H(u) = \ln V(u) \quad (18)$$

is also strictly increasing. Furthermore

$$H''(u) = \frac{V(u)V''(u)}{V^2(u)} - \frac{(V'(u))^2}{V^2(u)}$$

$$= \int_0^1 (w - V'(u)/V(u))^2 e^{uw} dw / V(u)$$

so  $H''(u) > 0$  for all real  $u$ , and  $H$  is convex.

Writing

$$\Delta_{\zeta}(p, x) = H(\zeta x(1-p)) - H(-xp), \quad (19)$$

(17) and (18) give

$$x_{\zeta}(p, x) = e^{\Delta_{\zeta}(p, x)}. \quad (20)$$

We must now distinguish two cases.

Case 1.  $\zeta \geq 1$ . For  $0 \leq p \leq 1$ ,

$$(\partial/\partial p)\Delta_{\zeta}(p, x) = -x[\zeta H'(\zeta x(1-p)) - H'(-xp)]. \quad (21)$$

Since  $H'(\zeta x(1-p)) \geq 0$ , and  $\zeta \geq 1$

$$\zeta H'(\zeta x(1-p)) - H'(-xp) \geq H'(\zeta x(1-p)) - H'(-xp)$$

$$\geq 0$$

since the convexity of  $H$  implies that  $H'$  is nondecreasing.

Equation (21) then yields

$$(\partial/\partial p)\Delta_{\zeta}(p, x) \leq 0,$$

so that

$$\Delta_{\zeta}(1, x) \leq \Delta_{\zeta}(p, x) \leq \Delta_{\zeta}(0, x)$$

for all  $0 < p < 1$ . But  $\Delta_{\zeta}(1, x) = -H(-x)$  and  $\Delta_{\zeta}(0, x) = H(\zeta x)$ , so

$$-H(-x) \leq \Delta_{\zeta}(p, x) \leq H(\zeta x)$$

or

$$\frac{1}{v(-x)} \leq f_{\zeta}(p, x) \leq v(\zeta x) \quad (22)$$

for all  $0 < p < 1$  and  $x > 0$ .

Case 2.  $\zeta \leq 1$ . In this case

$$\Delta_1(p, \zeta x) \leq \Delta_{\zeta}(p, x) \leq \Delta_1(p, x)$$

since  $H$  is nondecreasing. We saw in case 1, though, that  $\Delta_1(\cdot, x)$  is nonincreasing. Hence

$$\Delta_1(1, \zeta x) \leq \Delta_{\zeta}(p, x) \leq \Delta_1(0, x),$$

i.e.

$$-H(-\zeta x) \leq \Delta_{\zeta}(p, x) \leq H(x).$$

Therefore

$$\frac{1}{v(-\zeta x)} \leq f_{\zeta}(p, x) \leq v(x) \quad (23)$$

for all  $0 < p < 1$  and  $x > 0$ .

Returning for the moment to the general case, note that  $\lim_{u \rightarrow -\infty} V(u) = 0$ ,  $V(0) = 1$ , and  $\lim_{u \rightarrow \infty} V(u) = \infty$ , and recall that  $V$  is continuous and strictly increasing. Since  $\zeta\omega < 1$ , the equation

$$V(x') = 1/\zeta\omega \quad (24)$$

has a unique root  $x' = x'(\omega, \zeta)$  in  $(0, \infty)$ , while the equation

$$V(x'') = \zeta\omega \quad (25)$$

has a unique root  $x'' = x''(\omega, \zeta)$  in  $(-\infty, 0)$ . Now consider again the cases discussed above. In case 1 ( $\zeta \geq 1$ ) let

$$y = -x'' \text{ and } z = x'/\zeta. \quad (26)$$

Then from (22),

$$\frac{1}{V(x'')} = \frac{1}{V(-y)} \leq f_{\zeta}(p, v)$$

while

$$f_{\zeta}(p, z) \leq V(\zeta z) = V(x')$$

for all  $0 < p < 1$ . In case 2 ( $\zeta \leq 1$ ) let

$$y = -x''/\zeta \text{ and } z = x'. \quad (27)$$

Then from (23)

$$\frac{1}{V(x'')} \leq f_{\zeta}(p, y) \text{ and } f_{\zeta}(p, z) \leq V(x')$$

for all  $0 < p < 1$ . Therefore, in either case,

$$\frac{1}{\zeta\omega} \leq f_{\zeta}(p, y) \text{ and } f_{\zeta}(p, z) \leq \frac{1}{\zeta\omega}$$

for all  $0 < p < 1$ . Referring back to the sentence containing equation (17) we see that  $\psi_{y,\theta}$  is subregular and  $\psi_{z,\theta}$  is superregular for all  $0 < \theta \leq \min(1, 1/\zeta)$ . Q.E.D.

Though we will not need them below, we note that the proof gives simple formulas for  $y$  and  $z$ . The function  $V$  is defined by (16). The points  $x'$  and  $x''$  are defined in terms of  $V$  by (24) and (25), and  $y$  and  $z$  are defined in terms of  $x'$  and  $x''$  by (26) when  $\zeta \geq 1$  and by (27) when  $\zeta \leq 1$ .

It is easy to see that the classes of superregular and subregular functions are closed under addition and multiplication by nonnegative constants. Further, the constant functions are regular, hence both superregular and subregular. For any  $x > 0$ ,  $\psi_{x,\theta}(1) > \psi_{x,\theta}(0) = 1$ , therefore

$$\phi_{x,\theta}(p) = \frac{\psi_{x,\theta}(p) - 1}{\psi_{x,\theta}(1) - 1} \tag{28}$$

is superregular or subregular if  $\psi_{x,\theta}$  is. Also  $\phi_{x,\theta} \in D$ , with

$\phi_{x,\theta}(0) = 0$  and  $\phi_{x,\theta}(1) = 1$ . Thus, combining Lemma 9 and Lemma 5 we obtain the following theorem.

**Theorem 4.** There are positive constants  $\gamma = \gamma(\omega, \zeta)$  and  $z = z(\omega, \zeta)$  such that

$$\phi_{y,\theta}(p) \leq \gamma(p; \theta_1, \theta_2) \leq \phi_{z,\theta}(p) \quad (29)$$

for all  $0 < \theta \leq \min(1, 1/\zeta)$  and  $0 \leq p \leq 1$ .

**Corollary.** For any  $0 < \theta \leq \min(1, 1/\zeta)$  and  $0 \leq p \leq 1$

$$\gamma(p) \leq 1/e^{z(1-p)/\theta}, \quad (30)$$

so that

$$\lim_{\theta \rightarrow 0} \gamma(p) = 0 \quad (31)$$

if  $0 \leq p < 1$ .

**Proof.** A simple calculation shows that

$$\phi_{z,\theta}(p) = \frac{1}{e^{z(1-p)/\theta}} \frac{1 - e^{-zp/\theta}}{1 - e^{-z/\theta}},$$

and the second factor on the right clearly does not exceed unity.

Q.E.D.

Equation (30) suggests that when the learning rates are small the probability of being absorbed on the unfavorable side is very small.

Combining (31) with Theorem 2 we immediately obtain

Theorem 5. For  $0 < p < 1$

$$\chi(p) \sim \frac{2p}{\theta_2 \pi_2 - \theta_1 \pi_1}$$

as  $\theta \rightarrow 0$ .

Thus when  $\theta_2 \pi_2 > \theta_1 \pi_1$  (or, more generally, when  $\theta_2 \pi_2 \neq \theta_1 \pi_1$ ) and  $\theta$  is small, the mean number of alternations is of the order of magnitude of  $1/\theta$ , and tends to be inversely proportional to the difference  $|\theta_2 \pi_2 - \theta_1 \pi_1|$  in favorability between the two responses. When  $\theta_2 \pi_2 = \theta_1 \pi_1$  and  $\theta$  is small the mean number of alternations is of the order of magnitude of  $1/\theta^2$  by Theorem 3. Thus if the learning rates are small we expect many more alternations when the two responses are approximately equally favorable than when one is much more favorable than the other.

### 5. $\gamma$ as an Analytic Function

In this section we assume that the reader is familiar with the elements of the theory of analytic functions as presented, for example, by Knopp (1945), whose terminology we follow.

In order to motivate our results, consider the case  $\theta_2 = 1$ ,  $\theta_1 < 1$ . Then, since  $\gamma(0) = 0$ , (8) reduces to

$$\gamma(p) = \gamma(g_1(p)) \pi_1 p / [\pi_1 p + \pi_2 (1-p)] \quad (32)$$

where

$$g_1(p) = (1-\theta_1)p + \theta_1.$$

(We will also use the notation

$$g_2(p) = (1-\theta_2)p$$

below.) Iterating (32), and recalling that  $\gamma$  is continuous with  $\gamma(1) = 1$ , we obtain

$$\gamma(p) = \Lambda(p)/J(p) \tag{33}$$

where

$$\Lambda(p) = \prod_{n=0}^{\infty} g_{1,n}(p),$$

$$J(p) = \prod_{n=0}^{\infty} [g_{1,n}(p) + \frac{\pi_2}{\pi_1} (1-g_{1,n}(p))],$$

$g_{1,0}(p) = p$ , and  $g_{1,n}$  is the  $n^{\text{th}}$  iterate of  $g_1$ ,  $n \geq 1$ . Since  $g_{1,n}(p) = 1 - (1-\theta_1)^n(1-p)$ ,  $n \geq 0$ , the infinite products  $\Lambda(p)$  and  $J(p)$  converge for all complex numbers  $p$ . So the formula (33) serves to continue  $\gamma$  analytically into the complex plane  $C$ . If  $\pi_2 = \pi_1$  then  $J(p) = 1$  and  $\gamma$  is entire. If  $\pi_2 \neq \pi_1$  then  $\gamma$  has a pole of order 1 at each point that is a zero of one of the factors of  $J(p)$ . Since  $\pi_1 p + \pi_2(1-p) = 0$  if and only if  $p = c$  where

$$c = \pi_2/(\pi_2 - \pi_1),$$

these zeros are just the points  $g_{1,n}^{-1}(c)$  that  $g_{1,n}$  maps into  $c$  for some  $n \neq 0$ . Note that  $c > 1$  or  $c < 0$  depending on whether  $\pi_2 > \pi_1$  or  $\pi_2 < \pi_1$ , and that the sequence  $c, g_1(c), g_{1,2}(c), \dots$  of poles is confined to  $[c, \infty)$  or  $(-\infty, c]$ , respectively, in these two cases.

These examples and that given by Theorem 1 suggest all of the possibilities for the qualitative analytic character of  $\gamma$  that arise in the general case treated in Theorem 6, which summarizes our results. The theorem shows that  $\gamma$  is always meromorphic and occasionally entire. Whenever there are poles,  $c$  is the one closest to  $[0, 1]$ .

If  $c$  is a pole and its distance from  $[0, 1]$  is less than 1 then there will be points  $x$  in  $[0, 1]$  such that the Taylor series

$$\gamma(p) = \sum_{n=0}^{\infty} \frac{\gamma^{(n)}(x)}{n!} (p-x)^n \quad (34)$$

about  $x$  does not converge for all  $0 \leq p \leq 1$ . In such cases an attempt to compute the sequence  $\{\gamma^{(n)}(x)\}$  by the conventional method of substituting (34) into (8) and equating coefficients of  $(p-x)^n$  on the two sides of the resulting equation seems doomed to failure. In fact, little progress has been made to date in computing these coefficients even in cases where this obstacle is not present. Theorem 6 gives some information about them. For instance, when  $c$  is a pole the standard formula for the radius of convergence yields

$$\limsup_{n \rightarrow \infty} \sqrt[n]{|\gamma^{(n)}(x)|/n!} = \frac{1}{|x-c|}$$

for  $0 \leq x \leq 1$ . Moreover the theorem specifies the sign of  $\gamma^{(n)}(x)$  for all  $n \geq 0$ .

To state Theorem 6, some new notation will be needed. The function  $\Delta$  is defined by

$$\Delta(y) = \pi_2(1 - (1-\theta_2)^y) - \pi_1(1 - (1-\theta_1)^y) \quad (35)$$

for  $y \geq 1$ .  $E$  is the set of all points that map into  $c$  under repeated application of  $g_1$  and  $g_2$ , i.e.

$$E = \{g_{i_1}^{-1}(g_{i_2}^{-1}(\dots(g_{i_n}^{-1}(c)))) : n \geq 1 \text{ and } i_j = 1 \text{ or } 2 \text{ for all } 1 \leq j \leq n\} \cup \{c\}. \quad (36)$$

For any real  $y$ ,  $[y]$  is the smallest integer greater than  $y$ .

Finally, we note from (5) that  $U$  may be regarded as an operator on complex valued functions of a complex variable. We shall sometimes so regard it below.

Theorem 6. Suppose  $\theta_2 \pi_2 > \theta_1 \pi_1$  and  $1 > \theta_1, \theta_2$ .

a. If  $\pi_2 > \pi_1$  then  $\gamma$  can be continued analytically throughout  $C - E$ . The point  $c$  is a pole of order 1 and every other point of  $E$  is either a pole of order 1 or a regular point of  $\gamma$ . For all  $n \geq 1$  and  $0 \leq p \leq 1$ ,  $\gamma^{(n)}(p) > 0$ .

b. If  $\pi_2 = \pi_1$  then  $\gamma$  can be continued analytically over the entire complex plane. For all  $n \geq 1$  and  $0 \leq p \leq 1$ ,

$$0 < \gamma^{(n)}(p) \leq n! 2^n \frac{n(n-1)/2}{\prod_{j=1}^n (1-\eta_j)}, \quad (37)$$

where  $\eta = 1 - \min(\frac{1}{2}, \frac{1}{2})$ .

c. If  $\pi_2 < \pi_1$  then there is a unique  $x > 1$  for which  $\Delta(x) = 0$ .

i. If  $x$  is an integer, then  $\gamma$  is a polynomial of degree  $x$ , and  $\gamma^{(n)}(p) > 0$  for all  $1 \leq n \leq x$  and  $0 \leq p \leq 1$ .

ii. If  $x$  is not an integer then  $\gamma$  can be continued analytically throughout  $C - E$ . The point  $c$  is a pole of order 1, and every point of  $E$  is either a pole of order 1 or a regular point of  $\gamma$ . For any  $0 \leq p \leq 1$ ,  $\gamma^{(n)}(p) > 0$  if  $1 \leq n \leq [x]$  or if  $n - [x]$  is a positive even integer, and  $\gamma^{(n)}(p) < 0$  if  $n - [x]$  is a positive odd integer.

The functional equation (8) holds for all of the extensions described above.

The proof of Theorem 6 is fairly routine, but rather intricate. Limitations of space prevent inclusion of more than the following brief sketch.

Sketch of proof. An inequality is obtained relating the quantities  $|(U^m \psi)^{(n)}|$  for a smooth function  $\psi$ . When  $\psi(0) = 0$  and  $\psi(1) = 1$  this inequality yields a bound for  $|\gamma^{(n)}|$  on taking the limit as  $m \rightarrow \infty$ . When  $\pi_1 = \pi_2$  this bound is the expression on the right side of (37), and it follows that  $\gamma$  is entire. When  $\pi_1 \neq \pi_2$  this bound shows that  $\gamma$  can be extended to a function analytic in an oval including  $[0,1]$  and having  $c$  on its boundary. The functional equation (8) can be rewritten  $W\gamma = \gamma$  where

$$W\psi(z) = \psi(g_1(z))w(z) + \psi(g_2(z))(1-w(z))$$

and  $w(z) = \pi_1 z / (\pi_1 z + \pi_2(1-z))$ . The functions  $W^n \gamma$  provide a sequence of analytic continuations of  $\gamma$  into regions whose union is  $C - E$ . From the equation  $W\gamma(z) = \gamma(z)$  it can be seen that  $c$  is at worst a pole of order 1 of  $\gamma$ , and from this it follows that each point of  $E$  is either a pole of order 1 or a regular point. If  $c$  is regular, all are regular. To prove the assertion about the signs of the derivatives of  $\gamma$ , we consider, for each hypothesis about the parameters of the model, the class of nonnegative functions having derivatives with the pattern of signs indicated by the theorem, with, however, strict inequalities replaced by weak inequalities. It is shown that  $U$  preserves this class. Since  $I$  belongs to this class and  $(U^n I)^{(n)}(p) \rightarrow \gamma^{(n)}(p)$  as  $n \rightarrow \infty$ ,  $\gamma$  belongs to this class also. A supplementary argument using the analyticity of  $\gamma$  yields strict inequalities. Finally, the above results on the signs of the derivatives of  $\gamma$  and the entirety of  $\gamma$  if  $c$  is a regular point are shown to preclude regularity of  $c$  in cases (a) and (cii). Q.E.D.

Surveying the cases treated in Theorem 6 we see that in all of them  $\gamma^*(p) > 0$  for  $0 \leq p \leq 1$ . This implies that if  $p_1$  has a distribution  $F$  with mean  $\mu$  and positive variance, and if corresponding probabilities and expectations are indicated by subscript  $F$ 's, then

$$P_F(\liminf A_{1,n}) = E_F[\gamma(p_1)]$$

$$\geq \gamma(\mu) + \min_{0 \leq p \leq 1} \gamma''(p) E_F[(p_1 - \mu)^2]/2$$

$$> \gamma(\mu) = P_\mu(\liminf A_{1,n}).$$

Thus, if the distribution of  $p_1$  over a group of animals has mean  $1/2$  and positive variance, the proportion absorbed on the unfavorable side will tend to exceed the corresponding proportion for zero variance. ¶ Since all of the stat rats for the .75 group of Experiment 1 of Weinstein, et al (1965) had  $p_1 = 1/2$ , and since there is no reason to believe that this condition was met by all of the real rats, the above result may help to explain why a few more real rats than stat rats were absorbed on the unfavorable side.

## References

- Behrend, E. R., and Bitterman, M. E. Probability-matching in the fish. American Journal of Psychology, 1961, 74, 542-551.
- Bitterman, M. E. Phyletic differences in learning. American Psychologist, 1965, 20, 396-410.
- Bitterman, M. E., Wodinsky, J., and Candland, D. K. Some comparative psychology. American Journal of Psychology, 1958, 71, 94-110.
- Brody, A. L. Nonreinforcement in a noncorrection T maze. Journal of Comparative and Physiological Psychology, 1965, 60, 428-431.
- Bush, R. R., and Mosteller, F. Stochastic models for learning. New York: Wiley, 1955.
- Bush, R. R. Sequential properties of linear models. In R. R. Bush<sup>^</sup> and W. K. Estes (Eds.), Studies in Mathematical Learning Theory. Stanford: Stanford Univer. Press, 1959. Pp. 215-227.
- Karlin, S. Some random walks arising in learning models I. Pacific Journal of Mathematics, 1953, 3, 725-756.
- Kemeny, J. G., Snell, J. L., and Knapp, A. W. Denumerable markov chains. Princeton, N. J.: Van Nostrand, 1966.
- Knopp, K. Theory of functions, parts I and II. New York: Dover, 1945 and 1947.
- Meyer, D. R. The effects of differential probabilities of reinforcement on discrimination learning by monkeys. Journal of Comparative and Physiological Psychology, 1960, 53, 173-175.

- Mosteller, F., and Tatsuoka, M. Ultimate choice between two attractive goals: predictions from a model. Psychometrika, 1960, 25, 1-18.
- Norman, M. F. Some convergence theorems for stochastic learning models with distance diminishing operators. Journal of Mathematical Psychology, 1967, 4, in press.
- Parducci, A., and Polt, J. Correction vs. noncorrection with changing reinforcement schedules. Journal of Comparative and Physiological Psychology, 8, 51, 492-495.
- Stanley, J. C., Jr. The differential effects of partial and continuous reward upon the acquisition and elimination of a runway response in a two-choice situation. Ed. D. Thesis, Harvard University, 1950.
- Weinstock, S., North, A. J., Brody, A. L., and LeGuidice, Joann. Probability learning in a T maze with noncorrection. Journal of Comparative and Physiological Psychology, 1965, 60, 76-81.

**MEASUREMENT AND PSYCHOPHYSICS**

by

**DAVID KRANTZ**

at the

**American Mathematical Society Summer Seminar**

on the

**Mathematics of the Decision Sciences**

**Stanford University**

**July - August 1967**

## A SURVEY OF MEASUREMENT THEORY

### Table of Contents

1. Examples of measurement
  - 1.1 Definition of measurement
  - 1.2 Hölder's Theorem
  - 1.3 Extensive measurement
  - 1.4 Conjoint measurement
2. Some new problems generated by applying theories of measurement in the social sciences
  - 2.1 The role of measurement theory in social science
  - 2.2 Foundations of geometry
  - 2.3 Ordered rings
  - 2.4 Factorial designs; theory of models
  - 2.5 Semiorders
  - 2.6 Error theory
3. Metrics with additive segments
  - 3.1 Preview
  - 3.2 A bounded version of Hölder's Theorem
  - 3.3 The ternary relation  $\langle xyz \rangle$
  - 3.4 Existence and uniqueness of a metric
  - 3.5 Existence of segments
4. Polynomial measurement
  - 4.1 Independence and sign-dependence

4.2 Additive conjoint measurement and independent dimensions in geometry

4.3 Simple polynomials

5. The measurement of color

5.1 Metameric matches and vectorial representation

5.2 Photopigments

5.3 Color appearance

**This work was written for use at the American Mathematical Society Summer Seminar in Applied Mathematics to be held at Stanford, 1967. The work was partially supported by Public Health Service Grant GM-1231.**

## 1. Examples of measurement theory

### 1.1 Definition of measurement

Measurement, in its broadest sense, consists of the correspondence between mathematical objects, such as real numbers, vectors, or operators, and empirical objects, such as heavy bodies, forces, colors, etc. The correspondence is based on an isomorphism between observable formal properties of the empirical objects and the formal properties characterizing the mathematical objects. For example, in the measurement of mass, positive real numbers are assigned to heavy objects, so that the order of the numbers reflects the order of the objects, as determined by a suitable balance, and the addition of real numbers corresponds to combining of objects.

Many instances of measurement are like the measurement of mass, insofar as they involve construction of a real-valued function that preserves the order and additive structure of an empirical system. Such constructions are based ultimately on the theorem, due to Hölder [13], that any Archimedean fully ordered group is isomorphic to a subgroup of the ordered group of additive real numbers. I shall present a formal statement and proof of this theorem in the next section. Following this, I shall present two applications of Hölder's theorem. In the first application, the additive structure in the empirical objects is given directly, similar to combining heavy objects in the same pan of a balance. This is called *extensive measurement*. In the second application, no additive structure is given directly, but nevertheless, an associative binary operation can be defined, and Hölder's theorem applied.

## 1.2 Hölder's Theorem

**DEFINITION 1.1** Let  $G$  be a group, with binary operation  $(x,y) \rightarrow xy$  and identity  $e$ , and let  $\geq$  be a total order on  $G$ . The pair  $(G, \geq)$  is called an *ordered group* if for all  $x, y, z \in G$ ,  $x \geq y$  implies both  $xz \geq yz$  and  $zx \geq zy$ . The ordered group  $(G, \geq)$  is called *Archimedean* if for all  $x, y \in G$ , with  $x > e$ , there exists some positive integer  $n$  such that  $x^n > y$ .

We shall denote the ordered additive group of real numbers by  $(\mathbb{R}, +, \geq)$ .

**THEOREM 1.1** Let  $(G, \geq)$  be an Archimedean ordered group. Then  $(G, \geq)$  is isomorphic to a subgroup of  $(\mathbb{R}, +, \geq)$ . Moreover, the isomorphism is unique up to multiplication by a positive constant.

*Proof:* Let  $G^+ = \{x \mid x > e\}$ . We can distinguish 2 cases:

- A)  $G^+$  has a lower bound  $x_1 > e$ ;
- B)  $\inf G^+ = e$ . ( $G = \{e\}$  is a trivial case.)

In case A), for any  $y \in G$ , there exists a unique integer  $n$  (positive, negative, or zero) such that  $x_1^n \leq y < x_1^{n+1}$ . If  $y \neq x_1^n$ , then  $x_1^{-n}y$  is in  $G^+$  but is  $< x_1$ , a contradiction. Thus,  $y = x_1^n$ . Hence,  $G$  is cyclic with generator  $x_1$ , and the theorem follows; the subgroup of  $(\mathbb{R}, +, \geq)$  is any discrete subgroup.

For case B), let  $x \in G^+$  and  $y \in G$  be arbitrary; then there exists a unique integer  $N(x, y)$  such that  $x^{N(x, y)} \leq y < x^{N(x, y)+1}$ . Clearly, for  $x, x' \in G^+$ ,  $y \in G$ , we have

$$[N(x, x') + 1][N(x', y) + 1] > N(x, y) \geq N(x, x')N(x', y). \quad (1)$$

Let  $\{x_k\}$  be any sequence in  $G^+$  which converges to  $e$ . It is easy to show that for any  $y \in G^+$ ,  $N(x_k, y) \rightarrow +\infty$ , while for  $y^{-1} \in G^+$ ,  $N(x_k, y) \rightarrow -\infty$ . For any  $y, z \in G$ , with  $z \neq e$  and for  $k, \ell$  sufficiently large, we have by (1),

$$\frac{N(x_k, y)}{N(x_k, z)} < \frac{[N(x_k, x_\ell) + 1][N(x_\ell, y) + 1]}{N(x_k, x_\ell)N(x_\ell, z)}. \quad (2)$$

If we fix  $\ell$  and let  $k \rightarrow \infty$ , taking the limsup on the left and right in (2), we obtain

$$\limsup_{k \rightarrow \infty} \frac{N(x_k, y)}{N(x_k, z)} \leq \frac{N(x_\ell, y) + 1}{N(x_\ell, z)}. \quad (3)$$

Now taking the liminf on the right in (3), as  $\ell \rightarrow \infty$ , we find that  $\lim N(x_k, y)/N(x_k, z)$  exists (and is finite, as is easily seen). For fixed  $y_1 \in G^+$ , define

$$\phi(y) = \lim_{k \rightarrow \infty} \frac{N(x_k, y)}{N(x_k, y_1)}.$$

It is easily shown that  $\phi$  is an isomorphism of  $(G, \geq)$  onto a subgroup of  $(\mathbb{R}, +, \geq)$ . To this end, one can use the fact that for any  $x \in G^+$ ,  $y, z \in G$ ,

$$N(x, y) + N(x, z) + 1 \geq N(x, yz) \geq N(x, y) + N(x, z). \quad (4)$$

To show uniqueness of the isomorphism, let  $\phi'$  be any other isomorphism; then clearly, for any  $k, y$

$$N(x_k, y)\phi'(x_k) \leq \phi'(y) < [N(x_k, y) + 1]\phi'(x_k).$$

It follows that

$$\frac{\phi'(y)}{\phi'(y_1)} = \lim_{k \rightarrow \infty} \frac{N(x_k, y)}{N(x_k, y_1)} = \phi(y),$$

so that  $\phi' = \alpha\phi$ , where  $\alpha = \phi'(y_1) > 0$ . This completes the proof of Theorem 1.1.

For a different proof, see Birkhoff [3], p. 300. The proof given above has the advantages of being easily generalized, and of constructing the isomorphism  $\phi$  in a manner similar to actual measurement procedures. These points will be made more clearly in sections (1.3) and (2.2).

### 1.3 Extensive measurement

In extensive measurement, one starts with an empirical system that includes an associative binary operation. Placing 2 heavy objects together in the same pan of a balance is one example; others are found in the usual measurement procedures for length, where rods are combined by laying them end to end, and for time, where time intervals are concatenated by using the same event to mark the end of one interval and the beginning of another.

The following set of weak, logically independent axioms is due to Suppes [33].

*Primitives:*  $K$ , a nonempty set

$Q$ , a binary relation on  $K$

$\cdot$ , a binary function on  $K$ ,  $(x,y) \mapsto x \cdot y$ .

*Axioms:* For all  $x,y,z \in K$

1. if  $xQy$  and  $yQz$ , then  $xQz$
2.  $x \cdot y \in K$
3.  $(x \cdot y) \cdot z \ Q \ x \cdot (y \cdot z)$

4. if  $xQy$ , then  $x+z Q z+y$
5. if not  $xQy$ , then there exists  $w \in K$  such that  $x Q y+w$   
and  $y+w Q x$
6. not  $x+y Q x$
7. if  $xQy$ , then there is a positive integer  $n$  such that  $y Q nx$   
[ $1x = x$ ,  $nx = (n-1)x + x$ ].

We can prove the following measurement theorem.

THEOREM 1.2 If  $(K, Q, +)$  satisfies Axioms 1-7, then there exists a real-valued function  $\phi$  on  $K$  such that for all  $x, y \in K$

- (i)  $xQy$  if and only if  $\phi(x) \leq \phi(y)$
- (ii)  $\phi(x+y) = \phi(x) + \phi(y)$ .

Furthermore,  $\phi$  is unique up to multiplication by a positive constant.

Theorem 1.2 includes a representation theorem for extensive measurement--a theorem specifying that real-valued assignments can be constructed that preserve the empirically given structure--and a uniqueness theorem, limiting the class of possible representations. Uniqueness theorems are quite important in measurement, since they determine what sorts of statements about measured values are meaningful. Measurement representations that are unique up to multiplication by a positive constant are called *ratio scales*, because ratios are preserved by permissible changes in representation. Thus, the statement "X is twice as tall as Y" is meaningful, independent of the units chosen for measurement of length.

To prove Theorem 1.2, it is convenient to introduce a relation  $\sim$  on  $K$ :  
 $x \sim y$  if  $xQy$  and  $yQx$ . From the axioms,  $\sim$  can be shown to be an equivalence relation. Moreover, it can be shown that  $Q, +$  induce a total order,  $\geq$ , and a binary operation,  $+$ , on the set of equivalence classes,  $K/\sim$ . The system  $(K/\sim, +, \geq)$  satisfies all the properties of  $G^+$  in Hölder's Theorem; in particular, although inverses do not exist,  $K/\sim$  is closed under subtraction of smaller elements from larger ones (see Axiom 5). This permits the proof of Hölder's Theorem to be carried through, with no change, for  $(K/\sim, +, \geq)$ ; this latter is proved isomorphic to a subsemigroup of the additive semigroup of positive real numbers. The isomorphism yields the measurement representation required in Theorem 1.2. Uniqueness follows similarly from the uniqueness argument for Hölder's Theorem.

Finally, I should like to point out the close relation between the construction of the isomorphism  $\phi$ , in Theorem 1.1 or 1.2, and actual procedures for assigning real numbers to objects. Consider the case where  $K$  consists of straight rods, and  $x+y$  is formed by laying end-to-end one replica of rod  $x$  and one of rod  $y$ . Laying rods side-by-side permits comparisons, establishing  $xQy$ , etc. Measurement is carried out by forming a standard sequence  $x, 2x, \dots, nx$  (we use additive notation) laying  $1, 2, \dots, n$  replicas of  $x$  end-to-end. To measure  $y$  in feet, we form the ratio of  $N(x, y)$  to  $N(x, y_1)$ , where  $y_1$  is a standard foot-ruler. The generator of the standard sequence,  $x$ , is chosen sufficiently small to attain any desired accuracy of measurement. Equation (4) shows that the approximate measures,  $N(x, y)/N(x, y_1)$ , are approximately additive. The main point of Theorem 1.1 was to show that  $N(x, y)/N(x, y_1)$  converges as  $x$  is taken arbitrarily small.

#### 1.4 Conjoint measurement

In the social sciences, it is rare to find associative binary operations that can be used for extensive measurement. However, it is common to observe an ordering of objects, where position in the ordering depends on the values of 2 or more independently controllable factors. Such a situation is represented formally by a transitive relation  $\geq$  defined over a product set,

$A = \prod_{i=1}^n A_i$ . The simplest law governing the dependence of ordinal position

on the different factors is an additive one:

$$\phi(a_1, \dots, a_n) = \sum_{i=1}^n \phi_i(a_i)$$

where  $\phi$  is a real-valued, order-preserving function on  $A$  and each  $\phi_i$  is a real-valued function on  $A_i$ . If such functions can be constructed, we say that *additive conjoint measurement* is feasible for the system  $(A_1, \dots, A_n, \geq)$ . The functions  $\phi_i$  and  $\phi$  provide measurement scales for the factors  $A_i$  and the observed output, relative to which an additive law holds.

Additive conjoint measurement has a complex history. However, it was the publication of a set of sufficient conditions, by Luce and Tukey [24] in 1964, that created widespread interest. The most important anticipation of their work was published by Debreu [9]. The Luce-Tukey axioms, which apply to the case  $A = A_1 \times A_2$ , are essentially the following.

*Primitives:*  $A_1, A_2$ . nonempty sets

$\geq$ , a binary relation on  $A = A_1 \times A_2$ .

*Axioms:*

1.  $\geq$  is a weak order; i.e., it is transitive and any 2 elements of  $A$  are comparable.

2. Any change in one factor can be exactly compensated by a change in the other; i.e., if  $a \in A$ ,  $b_1 \in A_1$ , then there exists  $b_2 \in A_2$ , such that  $a \sim (b_1, b_2)$  ( $\sim$  means  $\geq$  and  $\leq$ ); and similarly for the other factor.

Axiom 2 is called the *solvability axiom*, since we "solve" for  $b_2$ , given  $a, b_1$ .

3. For any  $(a_1, a_2), (b_1, b_2), (c_1, c_2) \in A$ , if  $(a_1, b_2) \geq (b_1, c_2)$  and  $(b_1, a_2) \geq (c_1, b_2)$ , then  $(a_1, a_2) \geq (c_1, c_2)$ .

Axiom 3 is called the *cancellation axiom*, since, given an additive representation, we can add up the 2 antecedent inequalities and cancel  $b_1 + b_2$ , yielding the conclusion. This condition is illustrated geometrically in terms of indifference curves in the  $A_1 \times A_2$  plane, in Figure 1. In the theory of webs, this is called the Thomsen condition (see Aczél, Pickert, and Rado [1]).

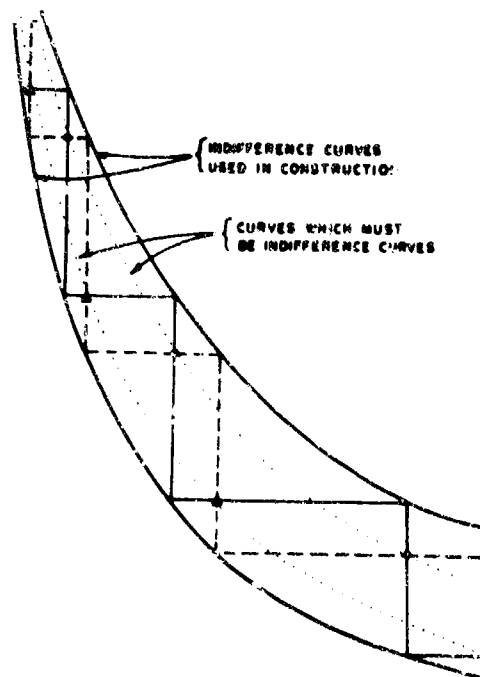


Fig. 1. The cancellation axiom illustrated for indifference curves. If two "flights of stairs" are inscribed between two indifference curves, as shown, then alternate intersections lie on the indifference curve when the cancellation axiom is true. (Taken from Luce & Tukey ([24], p.7)). (The author wishes to thank Academic Press, Inc. for permission to reprint this figure from the Journal of Mathematical Psychology.)

4. A sequence  $\{(a_{1i}, a_{2j}) \mid i, j = 0, \pm 1, \pm 2, \dots\}$  in  $A$  is called a *dual standard sequence* if  $(a_{1i}, a_{2j}) \sim (a_{1k}, a_{2\ell})$  iff  $i + j = k + \ell$ .  
If  $\{(a_{1i}, a_{2j})\}$  is a dual standard sequence, then for any  $a \in A$ , there exist  $n, m$  with  $(a_{1n}, a_{2n}) \geq a \geq (a_{1m}, a_{2m})$ .

Axiom 4 is called the *Archimedean axiom*. It is easily verified that Axioms 1, 3, and 4 are necessary for additive conjoint measurement; solvability is not.

THEOREM 1.3 If  $(A_1, A_2, \geq)$  satisfy Axioms 1-4, then there exist real-valued functions  $\phi$  on  $A$ ,  $\phi_1$  on  $A_1$ ,  $\phi_2$  on  $A_2$ , such that for all  $(a_1, a_2), (b_1, b_2) \in A$

- (i)  $(a_1, a_2) \geq (b_1, b_2)$  if and only if  $\phi(a_1, a_2) \geq \phi(b_1, b_2)$
- (ii)  $\phi(a_1, a_2) = \phi_1(a_1) + \phi_2(a_2)$ .

Furthermore, if  $\phi', \phi'_1, \phi'_2$  are any other such functions, then there are real numbers  $\alpha > 0, \beta_1, \beta_2$  such that  $\phi'_1 = \alpha\phi_1 + \beta_1, \phi' = \alpha\phi + \beta_1 + \beta_2$ .

It should be noted that the uniqueness clause of this theorem is the best that could be expected. Such a representation is called *interval scale measurement*; ratios of intervals are invariant under permissible transformations. (A more standard term would be *affine scale*, since the affine ratio is invariant.)

I shall sketch a proof of Theorem 1.3, based on Hölder's Theorem, which was published by Krantz [17].

Choose an arbitrary origin  $a^0 = (a_1^0, a_2^0)$  in  $A$ . By solvability, any equivalence class of  $A$  contains elements of forms  $(b_1, a_2^0)$  and  $(a_1^0, b_2)$ . Define an operation,  $+$ , on  $A/-$ , by

$$(b_1, a_2^0) + (a_1^0, b_2) = (b_1, b_2).$$

I shall show that  $\sim$  is well-defined and that  $(A/\sim, +, \geq)$  is an Archimedean ordered group, where  $\geq$  is defined in the natural way on  $A/\sim$ .

If  $(b_1, a_2^0) \sim (a_1^0, c_2)$  and  $(a_1^0, b_2) \sim (c_1, a_2^0)$ , then by cancellation,  $(b_1, b_2) \sim (c_1, c_2)$ . It follows that  $+$  is *well-defined* and *commutative*, since the equivalence class determined by adding arbitrary representations of the form  $(b_1, a_2^0)$ ,  $(a_1^0, b_2)$  is the same as that determined by adding arbitrary representations in the reverse order.

To prove *associativity* represent 3 arbitrary equivalence classes as  $(b_1, a_2^0)$ ,  $(a_1^0, b_2)$ ,  $(c_1, a_2^0)$ . By solvability find  $c_2, d_2 \in A_2$  such that  $(b_1, b_2) \sim (a_1^0, c_2)$  and  $(c_1, b_2) \sim (a_1^0, d_2)$ . By definition of  $+$ , cancellation, and commutativity,

$$\begin{aligned} [(b_1, a_2^0) + (a_1^0, b_2)] + (c_1, a_2^0) &= (c_1, c_2) \\ &= (b_1, d_2) \\ &= (b_1, a_2^0) + [(a_1^0, b_2) + (c_1, a_2^0)]. \end{aligned}$$

Obviously, the equivalence class of  $a^0$  is the *identity*, and if  $(b_1, b_2) \sim a^0$ , then  $(a_1^0, b_2)$  and  $(b_1, a_2^0)$  are *inverse*. Hence,  $(A/\sim, +)$  is a commutative group. Note that the results so far use only properties of  $\sim$ .

If  $(b_1, a_2^0) \geq (c_1, a_2^0)$ , and  $(a_1^0, d_2)$  is arbitrary, find  $d_1 \in A_1$  with  $(c_1, a_2^0) \sim (d_1, d_2)$ . By cancellation, applied to  $(b_1, a_2^0) \geq (d_1, d_2)$ ,  $(d_1, d_2) \sim (c_1, a_2^0)$ , we obtain  $(b_1, d_2) \geq (c_1, d_2)$ , or

$$(b_1, a_2^0) + (a_1^0, d_2) \geq (c_1, a_2^0) + (a_1^0, d_2).$$

Thus,  $(A/\sim, +, \geq)$  is an ordered group. Finally, the Archimedean property follows easily from Axiom 4.

By Hölder's Theorem, there is an isomorphism  $\phi$  of  $(A/ \sim, +, \succeq)$  onto a subgroup of  $(\mathbb{R}, +, \succeq)$ . Let  $\phi_1(b_1) = \phi(b_1, a_2^0)$ ,  $\phi_2(b_2) = \phi(a_1^0, b_2)$ . Then  $\phi(b_1, b_2) = \phi_1(b_1) + \phi_2(b_2)$  as specified by Theorem 1.3. The uniqueness clause follows from the fact that, if  $\phi'$ ,  $\phi'_1$ ,  $\phi'_2$  are any functions satisfying (i) and (ii) of Theorem 1.3, then

$$(b_1, b_2) \rightarrow \phi'_1(b_1) - \phi'_1(a_1^0) + \phi'_2(b_2) - \phi'_2(a_2^0)$$

is an isomorphism of  $(A/ \sim, +, \succeq)$  into  $(\mathbb{R}, +, \succeq)$ , and so, by Hölder's Theorem must differ from  $\phi$  by multiplication by a positive constant. This completes the proof.

Construction of the isomorphism  $\phi$ , and thus of the measurement scales  $\phi$ ,  $\phi_1$ ,  $\phi_2$ , depends on the construction of a standard sequence in  $(A/ \sim, +, \succeq)$ , as in the proofs of Theorems 1.1 and 1.2. This amounts to measuring the deviation of any  $(b_1, b_2)$  from  $(a_1^0, a_2^0)$  in terms of multiples of a small unit deviation from  $(a_1^0, a_2^0)$ .

2. Some new problems generated by applying theories of measurement in the social sciences.

2.1 The role of measurement theory in social science

For the physical scientist, measurement theory is properly a branch of philosophy. The axioms for extensive measurement of mass, length, or time provide a foundational analysis of long-established procedures. However, these axioms are too trivial to claim the status of laws of physics; rather, they are obvious properties of measurement operations, and are taken for granted in the actual practice of measurement.

Furthermore, in testing nontrivial laws that specify rules of combination for 2 or more variables, the physicist need not rely on an axiomatic analysis of the sort provided by conjoint measurement theory. For example, the equation of state for an ideal gas,  $pV/T = \text{constant}$ , and the second law of motion,  $F = ma$ , are stated in terms of numerical scales obtained by extensive measurement, and are directly testable by numerical calculations.

In the social sciences, there are no measurement procedures comparable to the ones used for measurement of mass, length, and time. Therefore, when an axiomatic theory of measurement is applied in a social science context, the axioms are not obvious properties of long-established procedures; rather, they are a set of proposed laws, which are not at all trivial. Some of the laws may be qualitative, i.e., directly testable by observations involving order or class membership. The axioms of additive conjoint measurement are of this sort. Other laws may be numerical,

for example, the assertion that 2 variables combine additively. These numerical laws cannot be tested as simply as in physics, since no numerical scales are specified for the variables. Rather, they must be tested by searching for numerical scales that satisfy the laws, or by testing other laws that imply or are implied by the given numerical laws. In short, at the present stage of development of quantitative theory in social science, it is impossible to separate the search for interesting empirical laws from the discovery and refinement of measurement procedures.

As a consequence of the above situation, measurement theory is of more than philosophical interest for social science. By providing an axiomatic theory for various numerical laws, one proposes qualitative experiments that distinguish between laws, and techniques of measurement where none existed previously. One result of the more dynamic and integral role of measurement theory in social science is that the discoveries or the difficulties encountered in empirical studies constitute an important source of new mathematical problems.

The next 5 sections are devoted to an overview of 5 areas in which new mathematical problems have emerged from the requirements of social science quantification: foundations of geometry, ordered rings, theory of models, semiorders, and error theory. The problems in foundations of geometry and in ordered rings were generated by the attempt to axiomatize laws other than the simple additive combination of variables: geometric laws, and polynomial combination laws, respectively. These topics will be pursued in more depth in lectures 3 and 4. The problems in theory of models, semiorders, and error theory derive from difficulties in

realizing idealized primitives of measurement theory, such as total orderings, amid the doubts and errors of real data.

## 2.2 Foundations of geometry

Geometrical models are heavily used in social science as a basis for quantitative treatments of similarity or correlation. For example, suppose that one has a set of objects,  $A$ , and obtains some measure of the dissimilarity of any 2 objects in  $A$ . This measure gives rise to an order relation on  $A \times A$ . To represent the dissimilarity ordering by a geometric model, one tries to map the objects of  $A$  into a metric space, where the ordering of metric distances corresponds to the observed ordering of dissimilarities. In 1962, Shepard [31] published a practical method of computing a representation for a finite set  $A$ , in low-dimensional Euclidean space, which yields the best approximation (for a given dimension) to the dissimilarity ordering on  $A \times A$ . Since then, this sort of measurement has been widely practiced, with little concern over appropriate foundations.

In terms of measurement theory, the problem of foundations may be stated as follows: given a set  $A$ , an observable ordering  $\geq$  of the pairs of elements of  $A$ , and a class  $C$  of metric spaces (the desired geometric representation), what axioms (empirical laws) must be satisfied, in order for there to be a metric  $d$  on  $A$ , such that  $(A, d)$  is in class  $C$ , and such that  $(x, y) \geq (z, w)$  if and only if  $d(x, y) \geq d(z, w)$ ? From the viewpoint of the classical field of foundations of geometry, we are asking for an

axiomatization of geometries of class  $C$ , in terms of the undefined (primitive) notions of a set of points and a *quaternary relation* on the points.

The classical axiom systems for foundations of Euclidean geometry (see Blumenthal [4]) generally involve undefined notions of *point*, *congruence of point pairs* (a quaternary relation), and *collinear betweenness of point triples* (a ternary relation). Sometimes, *lines*, and *incidence* of points and lines are also taken as primitive. In the study of empirical similarity, the required primitives (incidence of a point on a line, or collinear betweenness) do not seem to arise in any natural way. Thus, the problem of developing geometric measurement theories for similarity generates new problems in foundations of geometry, that is, axiomatizing different forms of metric geometry in terms of a single quaternary relation. One such axiomatization will be presented in detail in lecture 3, and some further possibilities will be mentioned briefly in lecture 4.

### 2.3 Ordered rings

Another source of problems is found in the general theory of conjoint measurement. Given a set of factors,  $A_1, \dots, A_n$ , and an order relation  $\geq$  on  $A = \prod_{i=1}^n A_i$ , various laws of combination for the different factors can be considered, besides the additive law discussed in lecture 1. The following definition is quite general.

DEFINITION 2.1 Let  $A_1, \dots, A_n$  be nonempty sets, with  $\geq$  a binary relation on  $A = \prod_{i=1}^n A_i$ . Let  $f$  be a real-valued function of  $n$  real variables. We say that  $(A_1, \dots, A_n, \geq)$  is *decomposable relative to  $f$*  if there exist real-valued functions  $\phi, \phi_1, \dots, \phi_n$ , with  $\phi$  defined on  $A$  and  $\phi_i$  on  $A_i$ , such that for  $a = (a_1, \dots, a_n), b = (b_1, \dots, b_n) \in A$ ,

$$(i) \quad a \geq b \text{ if and only if } \phi(a) \geq \phi(b)$$

$$(ii) \quad \phi(a) = f[\phi_1(a_1), \dots, \phi_n(a_n)].$$

The function  $f$  gives the rule of combination for the variables;  $\phi, \phi_1, \dots, \phi_n$  give appropriate measurement scales. The problem of measurement theory is to specify axioms (empirical laws) that are necessary and/or sufficient for decomposability relative to a specified rule  $f$ . This problem becomes fairly tractable when the function  $f$  is a polynomial in  $n$  variables. Moreover, quite a few miniature theories have been proposed, which explicitly posit polynomial rules of combination for a set of factors.

For one simple illustration of polynomial combinations rules, consider the relationship of the evaluative (moral) connotation of combinations of quantitative adverbs with adjectives, as a function of the adverb and of the adjective. The overall moral connotation of a combination such as "slightly evil" is better described as a multiplicative, rather than an additive combination of "slightly" and "evil". To see this, note that "slightly evil" would be rated better than "very evil", while "slightly pleasant" would be rated worse than "very pleasant". These opposite orderings of "slightly" and "very" correspond to multiplying numerical

scale values of the modifiers by moral values of opposite sign for "evil" and "pleasant". Studies of moral connotation that include the above examples, and use a multiplicative combination rule, were carried out by Cliff [8]. Many other miniature theories involve mixtures of additive and multiplicative combinations, i.e., more general polynomials.

The basic tool for polynomial conjoint measurement is the ring analog of Hölder's theorem: an Archimedean ordered ring (with nontrivial multiplication) is isomorphic to a unique subring of the ordered ring of real numbers. (See Birkhoff [3], p. 398). This tool can be used in at least 2 ways. One procedure is analogous to the proof of Theorem 1.3 on additive conjoint measurement: one introduces ring operations,

$+$ ,  $\cdot$  directly into the set of equivalence classes,  $A/\sim$ . The definitions of  $+$  and  $\cdot$  depend on the hypothesized polynomial; the required axioms are those for which the system  $(A/\sim, +, \cdot, \geq)$  becomes an Archimedean ordered ring. A different strategy is to let each relation statement of form  $a \geq b$  correspond to a suitable polynomial inequality. Obtaining the required functions  $\phi_i$  is equivalent to solving a set of simultaneous polynomial inequalities. This leads to the study of partial orders on polynomial rings. In particular, the following question seems to be unsolved and of interest: for what classes of partially ordered rings is an extension possible to an Archimedean total order? This problem is discussed by Tversky [3a]; for some results on extensions of partial orders, see Fuchs [11].

## 2.4 Factorial designs; theory of models

Given a binary relation  $\geq$  on  $\prod_{i=1}^n A_i$ , a common experimental procedure is to sample a finite subset  $B_i$  of  $A_i$ ,  $i = 1, \dots, n$ , and to observe the ordering  $\geq$  only on  $\prod_{i=1}^n B_i$ . This is called a factorial design. Certain axioms of polynomial conjoint measurement theories, such as solvability and Archimedean axioms, are untestable in such an experiment. But even if the untested axioms are valid in the entire empirical system  $(A_1, \dots, A_n, \geq)$ , while the testable axioms are verified in  $(B_1, \dots, B_n, \geq)$ , it may still be false that  $(B_1, \dots, B_n, \geq)$  is decomposable relative to the polynomial combination rule in question. Testing this decomposability amounts to searching for a simultaneous solution to a finite set of polynomial inequalities, a problem which is computationally demanding and for which there seems to be no general algorithm.

Thus, the problem arises of axiomatizing polynomial combination rules for finite systems. Here, the theory of models, developed by Tarski [34], is relevant. Using results from this theory, Scott and Suppes [30] proved a theorem that implies that there is no finite axiomatization for finite systems of additive conjoint measurement, by universal sentences in the first order functional calculus. One may conjecture that there is no finite axiomatization, in first-order functional calculus, for any system of polynomial conjoint measurement.

## 2.5 Semiorders

The binary or quaternary relations of extensive, conjoint, or metric space measurement are generally assumed to be transitive. In practice, 2 types of intransitivity are observed:

- (i)  $x \sim y, y \sim z$ , but  $x > z$ ,
- (ii)  $x > y, y > z$ , but  $z > x$ .

The first type may occur because differences between  $x$  and  $y$  and between  $y$  and  $z$  are too small to be detected, but add up to a detectable difference between  $x$  and  $z$ .

Type (i) intransitivities occur in a formal system called a *semi-order*, introduced by Luce [21]. This involves 2 binary relations,  $P$  (strict preference) and  $I$  (intransitive indifference).

Let juxtaposition denote the usual relation product, i.e.,  $x P I z$  if there exists  $y$  such that  $x P y$  and  $y I z$ ; let  $P^*$  be the reflection of  $P$  in the diagonal, i.e.,  $x P^* y$  if  $y P x$ . We can define a *semiorder* as follows.

DEFINITION 2.2  $(X, P, I)$  is a *semiorder* if  $X$  is a set, and  $P, I$  are binary relations on  $X$ , such that

- 1.  $(P, P^*, I)$  is a partition of  $X \times X$
- 2.  $P I P \subseteq P$
- 3.  $P^2 \cap I^2$  is empty.



Fig. 2. Illustration of axioms 2 (left) and 3 (right) for a semiorder. The configuration on the left implies  $xPw$ . The configuration on the right is asserted to be impossible.

The content of conditions 2 and 3 is depicted in Figure 2. In the left diagram,  $xPy$ ,  $yIz$ ,  $zPw$ , and the conclusion is  $xPw$ . In the right hand diagram,  $xP^2z$  (via  $y$ ) and  $xI^2z$  (via  $w$ ); the assertion is that no such configuration exists.

One of the main results on semiorders is that any semiorder defines a natural complete order.

**THEOREM 2.1** Let  $(X, P, I)$  be a semiorder. Define  $x \sim y$  if for all  $z \in X$ ,  $xIz$  iff  $yIz$ , and  $x \geq y$  if neither  $yPx$  nor  $yIx$ . Then  $\sim$  is an equivalence relation and  $\geq$  induces a total order on the equivalence classes  $X/\sim$ .

The proof of this theorem is a useful exercise.

Several problems arise in connection with semiorders. One problem is to axiomatize various forms of measurement, replacing the usual order relation by a semiorder. This can be done in a trivial way using the defined total order of Theorem 2.1, but the real point is to show that, in a semiordered system, one can attain any desired accuracy of measurement from

appropriate finite sets of  $P, I$  observations. This has been done for extensive measurement by Krantz [18].

A second type of problem is to deal with type (ii) intransitivities. One way to account for these is by assuming shifts in dimensions that determine the decision. For example, in purchasing a new car, each additional accessory may seem worth the added cost, but the total cost of several may drive one back to the basic model. Below some threshold, the cost dimension is ignored; above, it is decisive. One might capture this by assuming 2 semiorders  $(X, P_1, I_1)$ ,  $(X, P_2, I_2)$  over the same base set  $X$ , and defining the "lexicographic product",  $P = P_1 \cup (I_1 \cap P_2)$ ,  $I = I_1 \cap I_2$ . That is,  $xPy$  if  $xP_1y$  (the first dimension is decisive) or if  $x(I_1 \cap P_2)y$ . Obviously,  $P$  need not be transitive. The interesting question is to characterize lexicographic products of semiorders: given a pair of relations,  $P, I$ , what properties guarantee the existence of  $P_1, I_1, P_2, I_2$  such that  $(X, P_1, I_1)$  and  $(X, P_2, I_2)$  are semiorders and  $(P, I)$  is their lexicographic product? In empirical terms, can one infer the latent dimensional structure from a pattern of intransitivities?

## 2.6 Error theory

One of the most serious bars to testing the axioms of various measurement theories is the presence of "random" error. One way to deal with this difficulty is to superimpose a probability model on the algebraic one. For example, in conjoint measurement, one might assume that a pair  $(a_1, a_2)$  corresponds to a Gaussian random variable with expectation  $M(a_1, a_2)$ ; one

might interpret  $(a_1, a_2) \geq (b_1, b_2)$  to mean that  $M(a_1, a_2) \geq M(b_1, b_2)$ . Axioms such as transitivity or cancellation [1, 3 of (1.4)] are testable statistical hypotheses.

A criticism of the conventional statistical approach is that, if the measurement axioms are satisfied, then the construction of measurement scales induces transformations of the random variables. It is at least as reasonable to assume that the transformed random variables, rather than the original ones, satisfy a tractable probabilistic model, but this greatly complicates the statistical treatment.

More generally, one may wish to deal with random error in a manner that is less arbitrary than assumption of a special probabilistic model. It may be desirable to incorporate error processes more directly into the system of primitives and axioms.

An extreme version of the incorporation of error processes into a measurement axiomatization is to base the measurement entirely on error. For example, given a family of real-valued random variables, one may seek a transformation of the real numbers such that the transformed random variables are identically distributed except for translations. If this can be done, the transformation in question provides a measurement scale that regularizes the error theory. Levine [20] explored this problem quite deeply; among other results, he showed that if such a transformation exists, for a family of 3 or more random variables, then it is unique up to changes of origin and unit, i.e., we have interval scale measurement. Thus, the error theory has run away with the measurement procedure--there is no longer any room for basing measurement on an extensive operation,

a geometric model, or a polynomial combination rule. Some intermediate manner of incorporating random error in the measurement process would seem desirable.

### 3. Metrics with additive segments

#### 3.1 Preview

A metric space  $(X,d)$  is a *metric with additive segments* if for any  $x,z \in X$ , there is an isometry  $f$  of the real interval  $[0,d(x,z)]$  into  $X$ , such that  $f(0) = x$ ,  $f(d(x,z)) = z$ . Most metric spaces studied in geometry are of this type: e.g., Riemannian spaces, or G-spaces (Busemann, [7]). In this lecture, I examine the foundations of metrics with additive segments, starting with an ordering of pairs. More precisely, given a set  $A$ , and an ordering  $\succeq$  on  $A \times A$  (or a mapping  $(x,y) \mapsto xy$  of  $A \times A$  onto a totally ordered set,  $(P,\succeq)$ ), what axioms guarantee the existence of an order-preserving real-valued function  $\phi$  on  $P$  such that for  $d(x,y) = \phi(xy)$ ,  $(A,d)$  is a metric with additive segments? The source of this problem is the demand for a geometric model of dissimilarities, discussed in (2.2).

The key to analyzing foundations of metrics with additive segments is the ternary relation  $\langle xyz \rangle$ , which, in terms of a metric, can be defined as  $d(x,y) + d(y,z) = d(x,z)$ . We must define this relation, and establish its main properties, using the ordering alone. Once this is done, we define a binary operation in  $P$  as follows:  $xy + x'y' = uv$  if  $xy = uv$ ,  $x'y' = vw$ , and  $\langle uvw \rangle$ . This operation, however, cannot necessarily be defined for all pairs  $(xy, x'y')$  (there may not exist additive segments of arbitrary length). Thus, in order to apply Hölder's theorem to the system  $(P, +, \succeq)$ , we must establish a version of it that applies when the

binary operation is defined only for sufficiently small elements. This sort of local theorem has other important applications. In the next section, I shall state it, sketch its proof, and indicate the applications to extensive and conjoint measurement. In succeeding sections, I shall return to the question of metrics with additive segments.

### 3.2 A bounded version of Hölder's Theorem

DEFINITION 3.1 Let  $G$  be a set, with binary relations  $\geq$ ,  $B$  on  $G$ , and a binary operation  $(x,y) \mapsto x + y$  from  $B$  to  $G$ . The quadruple  $(G, B, +, \geq)$  will be called a *positive ordered local semigroup* if the following are true for all  $x, y, z, x', y' \in G$ :

1.  $\geq$  is a total order
2. if  $(x, y) \in B$ ,  $x \geq x'$ ,  $y \geq y'$ , then  $(y', x') \in B$
3. if  $(x, y), (x+y, z) \in B$ , then  $(y, z), (x, y+z) \in B$  and  
 $(x + y) + z = x + (y + z)$
4. if  $x \geq y$  and  $(x, z) \in B$ , then  $x + z \geq y + z$  and  
 $z + x \geq z + y$
5. if  $(x, y) \in B$ , then  $x + y > x$
6. if  $z > x$ , then there exists  $y \in G$  with  $(x, y) \in B$  and  
 $z \geq x + y$ .

A positive ordered local semigroup is *Archimedean* if for all  $x, y \in G$ ,  $\{n | nx \text{ defined, } y \geq nx\}$  is finite.

Note that by property 2,  $(x, y) \in B$  iff  $(y, x) \in B$ ; from this, we know that  $y + z, z + x, z + y$  are defined in property 4.

**THEOREM 3.1** *Let  $(G, B, +, \geq)$  be an Archimedean positive ordered local semi-group. Let  $G' = \{x | \exists y, (x, y) \in B\}$ . Then there is a real-valued function  $\phi$  on  $G'$  such that for all  $x, y \in G'$*

- (i)  $x \geq y$  iff  $\phi(x) \geq \phi(y)$
- (ii) if  $(x, y) \in B$ , then  $\phi(x + y) = \phi(x) + \phi(y)$ .

*Moreover, if  $\phi, \phi'$  are any 2 such functions, then  $\phi' = \alpha\phi$  for some  $\alpha > 0$ .*

The proof of Theorem 3.1 is like that of Theorem 1.1 in all essential details. We note only 2 slight differences. First, for  $x, y \in G'$ ,  $y \geq x$ , define  $N(x, y)$  to be the largest  $n$  for which  $nx$  is defined and  $y \geq nx$ . If  $(x, y) \in B$ , then  $y \geq nx$  implies  $(n + 1)x$  is defined; hence, for  $(x, y) \in B$ , we have  $[N(x, y) + 1]x > y \geq N(x, y)x$ , as in Theorem 1.1.

Second, the use of inverses in Theorem 1.1 is solely to provide elements of form  $y^{-1}x$ , where  $y < x$ . The same effect is achieved here by finding  $y'$  such that  $y + y' \leq x$ , using property 6 of Definition 3.1.

Theorem 3.1 is clearly applicable to extensive measurement, for the case where there is a practical upper bound on the size of elements that can be compared. This has been discussed by Luce and Marley [23]. Less obvious is the application of the theorem to a more realistic version of conjoint measurement. The solvability axiom (Axiom 2 of (1.4)) essentially forces the set  $A/\sim$  to be a subgroup of real numbers, whereas in practice, one would like to restrict attention to a bounded subset of such a subgroup. This corresponds to the fact that one cannot in practice always solve

equations of the form  $a \sim (b_1, b_2)$  for  $b_2$ , given  $a, b_1$ . The change on the first factor may be so large, as to be unmatchable within  $A_2$ . To deal with this case, solvability has been replaced by a much more realistic assumption (Debreu [9]; Luce [22]) called *restricted solvability*:

- 2'. For all  $a \in A$ ,  $b_1 \in A_1$ , if there exist  $\underline{c}_2, \bar{c}_2$  such that  $(b_1, \bar{c}_2) \geq a \geq (b_1, \underline{c}_2)$ , then there exists  $c_2$  such that  $a \sim (b_1, c_2)$ ; and a similar assumption with the roles of  $A_1, A_2$  interchanged.

In this more restricted situation, one can order "positive differences" between elements of  $A_1$  by comparison with a "difference" in  $A_2$ : namely define

$$a_1 - b_1 \geq a'_1 - b'_1 \quad \text{if there exist } a_2, b_2 \in A_2 \text{ such that} \\ (a_1, b_2) \geq (b_1, a_2), \quad (b'_1, a_2) \geq (a'_1, b_2).$$

Certain positive differences can then be "added" by laying off equivalent differences end-to-end. An additional axiom, similar to the cancellation axiom of (1.4), is required, and minor modifications of the Archimedean axiom (Axiom 4 of (1.4)) are needed, but given these, the bounded Hölder's theorem can be applied to the system of positive differences on each factor,  $A_1$  and  $A_2$ . Ultimately, this leads to the same conclusion as that of Theorem 1.3, based on much weaker assumptions\*.

---

\* This use of positive differences on each component in a system of additive conjoint measurement is unpublished; it draws on material from a book in preparation, by R. D. Luce, P. Suppes, A. Tversky, and the present author. For a slightly different treatment based on Axiom 2', see Luce [22].

### 3.3 The ternary relation $\langle xyz \rangle$

We return now to consideration of a set  $A$ , and a mapping  $(x,y) \rightarrow xy$  of  $A \times A$  onto a total order  $(P, \geq)$ . If there exists a function  $\phi$  from  $P$  to the reals, which preserves order, such that  $d(x,y) = \phi(xy)$  is a metric with additive segments, then the following 4 axioms are easily seen to be necessary.

1. For  $x \neq y$ ,  $xx = yy < xy$ .
2.  $xy = yx$ .
3. If  $xy \leq uw$ , then there exists  $v$  such that  $xy = uv$  and  $\langle uvw \rangle$ .
4. If  $x \neq y$ , then for any  $u, w$  there exist  $x_0, \dots, x_n$  such that  $x_0 = u$ ,  $x_n = w$ , and for  $i = 1, \dots, n$ ,  $x_{i-1}x_i \leq xy$ .

Axioms 1, 2, and 4 are stated entirely in terms of the ordering  $\geq$ , but Axiom 3 involves the ternary relation  $\langle uvw \rangle$ . In terms of the desired metric, this means  $d(u,v) + d(v,w) = d(u,w)$ . However, we shall define  $\langle \rangle$  in terms of the ordering alone. We do this by noticing that, if  $d(x,y) + d(y,z) = d(x,z)$ , then the distance  $d(y',z)$  achieves a minimum at  $y' = y$ , for all points  $y'$  on or inside the sphere with center  $x$  and radius  $d(x,y)$ . This characterization of  $y$  uses only ordinal relations.

DEFINITION 3.2  $\langle xyz \rangle_L$  if for all  $x', y', z'$  such that  $x'y' \leq xy$  and  $xz \leq x'z'$ , both of the following hold:

- (i)  $yz \leq y'z'$
- (ii) if  $yz = y'z'$ , then  $xy = x'y'$  and  $xz = x'z'$ .

Define  $\langle xyz \rangle$  if both  $\langle xyz \rangle_L$  and  $\langle zyx \rangle_L$ .

Roughly speaking,  $\langle xyz \rangle_L$  holds if  $yz$  is minimal among all  $y'z'$  such that  $y'$  is inside a sphere of radius  $xy$  and  $x'$  is outside a concentric sphere with radius  $xz$ . The relation  $\langle xyz \rangle$  is simply the symmetric form: clearly,  $\langle xyz \rangle$  iff  $\langle zyx \rangle$ . Henceforth, in Axiom 3 above, the relation  $\langle \rangle$  will be understood to be the one defined by Definition 3.2. From this definition, we obtain the following useful lemma (only Axiom 2 is used in the proof).

LEMMA 3.1 If  $\langle xyz \rangle$ ,  $x'y' \leq xy$ ,  $y'z' \leq yz$ , and  $xz \leq x'z'$ , then  $\langle x'y'z' \rangle$ .

This follows because if any inequality were strict, then (ii) of the definition would yield a contradiction of one of the other inequalities. Hence  $x'y' = xy$ ,  $y'z' = yz$ , and  $xz = x'z'$ , and  $\langle x'y'z' \rangle$  follows. We also note that from Axioms 1 and 2 and Definition 3.2, if  $\langle xyz \rangle$ , then  $xy, yz \leq xz$ . Given these preliminary results, we can prove the following fundamental theorem.

THEOREM 3.2 If Axioms 1-3 hold, and if  $\langle xyz \rangle$  and  $\langle xzw \rangle$ , then  $\langle yzw \rangle$  and  $\langle xyw \rangle$ .

Proof: First we prove  $\langle yzw \rangle$ . From  $xz \leq xw$  and  $\langle xyz \rangle_L$ , we have  $yz \leq yw$ . By Axiom 3, choose  $z'$  such that  $yz = yz'$  and  $\langle yz'w \rangle$ . By Lemma 3.1, it suffices to show  $zw \leq z'w$ . Since  $\langle xyz \rangle_L$  and  $yz' \leq yz$ , we have  $xz' \leq xz$ ; but then, by  $\langle xzw \rangle_L$ ,  $zw \leq z'w$  follows as required.

Next we show  $\langle xyw \rangle$ . Note that  $wy \leq wx$ ; otherwise, if  $wx < wy$ , then  $\langle wzx \rangle_L$  implies  $zx < zy$ , contradicting  $\langle zyx \rangle$ . By Axiom 3, choose  $y'$  with  $wy = wy'$  and  $\langle wy'x \rangle$ . By Lemma 3.1, it will suffice to show that  $xy \leq xy'$ .

From  $\langle yzx \rangle$ ,  $wz \leq wy'$ . Construct  $z'$  with  $wz = wz'$  and  $\langle wz'y' \rangle$ . Suppose  $xy' < xy$ . Then by  $\langle xyz \rangle$ , either  $xz' < xz$ , or  $yz < y'z'$ . The former is false, since  $\langle xzw \rangle$  and  $xz' < xz$  imply  $zw < z'w$ ; and the latter is wrong, because  $\langle wz'y' \rangle$ ,  $wz = wz'$ , and  $wy = wy'$  imply  $y'z' \leq yz$ . Hence, we conclude that  $xy \leq xy'$ , as required. This completes the proof of Theorem 3.2

Theorem 3.2 states the basic property of the ternary relation  $\langle \rangle$ , which is needed to construct a metric; as will be seen in the next section, it corresponds to associativity of the operation  $+$  defined on  $P$ .

### 3.4 Existence and uniqueness of a metric

DEFINITION 3.3 Let  $A, P$ , be as above, and let  $\langle \rangle$  be given by Definition 3.2. Define a binary operation  $+$  on  $P$  by  $xy + x'y' = uw$  if  $xy = uv$ ,  $x'y' = vw$ , and  $\langle uvw \rangle$ .

We note that this is well defined. Let  $P_1 = P - \{xx\}$ , and let  $B = \{(xy, x'y') \mid xy + x'y' \text{ is defined}\}$  is defined. Then the following theorem can be proved.

**THEOREM 3.3\*** *If Axioms 1-3 of (3.3) hold, then  $(P_1, B, +, \geq)$  is a positive ordered local semigroup; if Axiom 4 also holds, then it is Archimedean as well. Hence, if Axioms 1-4 hold, then there is a metric  $d$  on  $A$  such that*

$$(i) \quad xy \geq x'y' \text{ iff } d(x,y) \geq d(x',y')$$

$$(ii) \quad \langle xyz \rangle \text{ iff } d(x,y) + d(y,z) = d(x,z).$$

*Moreover,  $d$  is unique up to similarity transformations (i.e.,  $d$  is a ratio scale).*

*Proof:* The proof that  $(P_1, B, +, \geq)$  is a positive ordered local semigroup if Axioms 1-4 hold is almost immediate. To illustrate, we prove associativity (in the sense of property 3, Definition 3.1). Suppose  $\langle xy, x'y' \rangle$  and  $\langle xy + x'y', x''y'' \rangle \in B$ . Let  $uv = xy + x'y'$  and  $vw = x''y''$ , with  $\langle uvw \rangle$ . Let  $u'v' = xy$ ,  $v'w' = x'y'$ , with  $\langle u'v'w' \rangle$ . Then  $u'w' = uv$ . Since  $u'v' \leq uv$ , there exists  $z$  with  $u'v' = uz$  and  $\langle uzv \rangle$ . By Theorem 3.2,  $\langle zvw \rangle$  and  $\langle uzv \rangle$ . By Definition 3.2,  $zv = x'y'$ . Thus,  $x'y' + x''y''$  is defined and  $= zw$ ; so  $xy + (x'y' + x''y'')$  is defined and  $= uw = (xy + x'y') + x''y''$ .

From Theorem 3.1, there exists a real-valued function  $\phi$  on  $P_1^1$  such that  $xy \geq x'y'$  iff  $\phi(xy) \geq \phi(x'y')$  and  $\phi(xy + x'y') = \phi(xy) + \phi(x'y')$ . We note that there is at most one  $p \in P_1 - P_1^1$ , i.e., a maximal element of  $P$ , if such exists. We define  $d$  on  $A \times A$  by

$$d(x,y) = \begin{cases} \phi(xy) & \text{if } xy \in P_1^1, \\ 0 & \text{if } x = y, \\ \sup\{\phi(uv) \mid uv \in P_1^1\} & \text{if } xy \text{ is maximal.} \end{cases}$$

\* The results in (3.3)-(3.5) are essentially due to Beals and Krantz [2]. They considered a somewhat more general situation. The version presented here, particularly Theorem 3.3, draws on unpublished material from a book in preparation by Luce, Suppes, Tversky, and Krantz.

Obviously,  $xy \leq x'y'$  iff  $d(x,y) \leq d(x',y')$  and  $\langle xyz \rangle$  iff  $d(x,y) + d(y,z) = d(x,z)$ . The triangle inequality follows from the definition of  $\langle \rangle$ . Also, by Axiom 4, if  $xy$  is maximal in  $P$ , then  $\sup\{\phi(uv) \mid uv \in P_1^1\}$  is finite. Thus,  $d$  is a metric satisfying (i) and (ii). Clearly, any other metric  $d'$  with the same properties defines a function  $\phi'$  on  $P_1^1$  with the same properties as  $\phi$ ;  $\phi' = \alpha\phi$ , hence,  $d' = \alpha d$ , follow from the uniqueness assertion of Theorem 3.1. This completes the proof of Theorem 3.3.

### 3.5 Existence of segments

Note that Axioms 1-4 of (3.3) can be satisfied by the set  $A = \{x,y,z\}$ , with  $xy = yz < xz$ . In fact,  $\langle xyz \rangle$  holds, and the only  $d$  satisfying Theorem 3.3 is given by  $d(x,y) = d(y,z) = D$ ,  $d(x,z) = 2D$ , where  $D > 0$  is arbitrary. Such finite examples are avoided if we impose the requirement that any  $x,z \in A$  be joined by an *additive segment*, as defined in (3.1). One way to guarantee this is to impose 2 additional conditions, nondiscreteness and completeness:

5.  $P - \{xx\}$  has no minimal element.
6. If  $x_i$  is a sequence in  $A$  such that for  $u \neq v$ ,  $x_i x_j \leq uv$  for all but finitely many  $(i,j)$ , then there exists  $y \in A$  such that for  $u \neq v$ ,  $x_i y \leq uv$  for all but finitely many  $i$ . That is, any Cauchy sequence converges.

We call a subset  $\gamma$  of  $A$  a *partial segment from  $x$  to  $z$*  if (i)  $x,z \in \gamma$  and (ii) for any  $u,v \in \gamma$ ,  $\langle xuv \rangle$  or  $\langle xvu \rangle$ . The set  $\gamma$  is a *segment from  $x$  to  $z$*  if it is a maximal partial segment from  $x$  to  $z$ . By Zorn's lemma,

for any  $x, z$ , there exists a segment  $\gamma$  from  $x$  to  $z$ . We wish to show that  $\gamma$  is isometric to  $[0, d(x, z)]$ . For  $y \in \gamma$ , let  $f(y) = d(x, y)$ . Obviously,  $d(u, v) = |f(u) - f(v)|$ , so  $f$  is an isometry of  $\gamma$  into  $[0, d(x, z)]$ . It remains only to show that  $f$  is onto. For this, we use Axioms 5 and 6 above.

Let  $t$  be  $\in (0, d(x, z))$ . Let  $\gamma_1 = \{y \in \gamma \mid f(y) \leq t\}$ ,  $\gamma_2 = \{y \in \gamma \mid f(y) \geq t\}$ . From Axiom 6 it is easily shown that  $f$  attains its maximum in  $\gamma_1$  at some  $y_1 \in \gamma_1$  and its minimum in  $\gamma$  at  $y_2 \in \gamma_2$ . For example, if  $u_i$  is a sequence in  $\gamma_1$  such that  $d(x, u_i) \rightarrow \sup\{f(y) \mid y \in \gamma_1\}$ , then by Axiom 6,  $u_i \rightarrow y_1 \in A$ , and it is easy to see that  $\gamma \cup \{y_1\}$  is a partial segment. It follows by maximality that  $y_1 \in \gamma$ ;  $y_2$  is treated similarly. If  $y_1 = y_2$ , then  $f(y_1) = f(y_2) = t$ , and the required preimage of  $t$  in  $\gamma$  has been constructed. But for  $y_1 \neq y_2$ , we can use Axioms 5 and 3 to choose  $y$  with  $\langle y_1, y, y_2 \rangle$ , and  $y \notin \gamma_1, \gamma_2$ . By construction,  $y \notin \gamma$ , but by Theorem 3.2,  $\gamma \cup \{y\}$  is a partial segment, contradicting the maximality of  $\gamma$ . Thus,  $y_1 = y_2$  as required. This completes the proof of the following theorem (whose converse is obviously true also):

**THEOREM 3.4** *Let  $A, P, \geq$  satisfy Axioms 1-6. Then there is a metric  $d$  on  $A$ , unique up to multiplication by a positive constant, such that  $(A, d)$  is a complete metric space with additive segments, and such that  $xy \geq x'y'$  iff  $d(x, y) \geq d(x', y')$ .*

#### 4. Polynomial measurement theories

##### 4.1 Independence and sign-dependence

If  $\geq$  is a binary relation on  $\prod_{i=1}^n A_i$ , where  $n \geq 2$ , it induces other binary relations on products of any  $m$  factors, where  $m < n$ . For example, if  $b$  is a fixed element of  $\prod_{i=m+1}^n A_i$ , and  $a, a'$  are elements of  $\prod_{i=1}^m A_i$ , we can define

$$a \geq(b) a' \text{ if } (a, b) \geq (a', b).$$

Thus any choice of a fixed  $b \in \prod_{i=m+1}^n A_i$  induces a relation  $\geq(b)$  on  $\prod_{i=1}^m A_i$ . Similarly, choosing fixed components in any subset of factors induces a binary relation over the product of the remaining factors.

One of the things that makes the study of polynomial combination laws fruitful is that, for binary relations obeying such laws, the induced relation  $\geq(b)$  varies in regular and interesting ways as a function of the vector of fixed components,  $b$ . Recall that, according to Definition 2.1,  $(A_1, \dots, A_n, \geq)$  satisfies a polynomial combination law  $f$  provided that there are real-valued functions  $\phi_i$  on  $A_i$  and  $\phi$  on  $\prod_{i=1}^n A_i$ , such that  $\phi$  is order-preserving and  $\phi = f(\phi_1, \dots, \phi_n)$ . A simple example of regularity of induced relations occurs if

$f(x_1, \dots, x_n) = g(x_1, \dots, x_m) + h(x_{m+1}, \dots, x_n)$ . In that case, it is obvious that  $\geq(b)$  is independent of  $b$ , for  $b \in \prod_{i=m+1}^n A_i$ ; the ordering of elements of  $\prod_{i=1}^m A_i$  depends only on the values of  $g(\phi_1, \dots, \phi_m)$ . We say in this case that  $\prod_{i=1}^m A_i$  is independent of  $\prod_{i=m+1}^n A_i$ . In the simplest case,  $f(x_1, \dots, x_n) = \sum_{i=1}^n x_i$ , and any subset of factors is independent of its complement. We say in this case that the system

$(A_1, \dots, A_n, \geq)$  is *completely independent*; complete independence is thus a necessary condition for additive conjoint measurement in  $n$  factors.

A subset of factors can be independent of a proper subset of its complement, even when it is not independent of the entire complement; this occurs when an induced relation,  $\geq(b, c)$ , does not depend on the vector of components represented by  $b$ , for fixed  $c$ . For example, when  $f(x_1, x_2, x_3) = (x_1 + x_2)x_3$ ,  $A_1$  is independent of  $A_2$ , and vice versa, although  $A_1$  need not be independent of  $A_2 \times A_3$ .

If  $f(x_1, \dots, x_n) = g(x_1, \dots, x_m) \cdot h(x_{m+1}, \dots, x_n)$ , then the induced ordering of elements in  $\prod_{i=1}^m A_i$  depends not only on  $g(\phi_1, \dots, \phi_m)$  but on whether  $h(\phi_{m+1}, \dots, \phi_n)$  is positive, negative, or zero. Thus,  $\prod_{i=m+1}^n A_i$  can be partitioned into at most 3 subsets,  $S^+, S^0, S^-$ . All induced orders  $\geq(b)$  are identical for  $b \in S^+$ , and are the reverse of  $\geq(b)$  for  $b \in S^-$ , while  $\geq(b)$  is degenerate (the universal relation) for  $b \in S^0$ . We express this partition property by saying that  $\prod_{i=1}^m A_i$  is *sign-dependent on*  $\prod_{i=m+1}^n A_i$ . Independence of is the special case of *sign-independence on* in which only  $S^+$  or  $S^-$  is nonempty. Thus, for  $(x_1 + x_2)x_3$ ,  $A_1 \times A_2$  and  $A_3$  are mutually sign-dependent. It is not true, however, that  $A_1 \times A_3$  is sign-dependent on  $A_2$ . In fact, for  $x'_3 > x_3$ ,

$$(x_1 + y_2)x_3 \geq (x'_1 + y_2)x'_3 \quad \text{iff} \quad y_2 \leq \frac{x_1x_3 - x'_1x'_3}{x'_3 - x_3}$$

so that the value of  $y_2 = \phi_2(a_2)$  at which the order of  $(x_1, x_3)$  and  $(x'_1, x'_3)$  reverses is not fixed, leading to a partition, but varies with  $x_1, x_3, x'_1, x'_3$ .

Independence and sign-dependence properties are good examples of qualitative laws. Independence is subject to straightforward experimental testing, wherever order relations can be determined; when satisfied, it strongly indicates that the effects of certain variables can be evaluated apart from consideration of the fixed values of other variables. Sign-dependence is similarly testable; moreover, it has a kind of special "flavor", since a large effect can be produced in 2 completely different ways, by combination of two "positive" values or of two "negative" values. An instance of sign-dependence was discussed in section (2.4) where moral evaluations of descriptive adverb-adjective combinations were studied, e.g., "slightly evil", etc. There, the adverb factor is sign-dependent on the adjective factor; "evil" is in the S-part of  $A_2$ . It scarcely needs experimental testing to show that sign-dependence is more appropriate than independence, and hence, that multiplication is more appropriate than addition. The detailed accuracy of a multiplicative model is, of course, another question.

In general, particular polynomials exhibit more or less idiosyncratic patterns of independence and sign-dependence; thus, it is possible, on the basis of empirical information concerning these properties, to diagnose appropriate polynomial combination rules, or at least, to narrow the field.

#### 4.2 Additive conjoint measurement and independent dimensions in geometry

It was indicated previously that a system of additive conjoint measurement is completely independent. Surprisingly, a partial converse can be

proved: if  $(A_1, \dots, A_n, \succeq)$  satisfies complete independence, restricted solvability, and an Archimedean condition, and if  $n \geq 3$ , then  $(A_1, \dots, A_n, \succeq)$  is decomposable relative to  $\sum_{i=1}^n x_i$ . (Recall that restricted solvability means that any shift on  $n - 1$  factors can be compensated by a shift on the  $n$ th factor, given that certain boundedness conditions hold.) Since restricted solvability and Archimedean conditions are usually assumed to be valid empirically, this means that if  $n \geq 3$  (in the sense that there are at least 3 nontrivial factors), complete independence is the empirical equivalent to additivity of the factors.\* (No such result holds for  $n = 2$ ; in that case, the cancellation axiom (Axiom 3 of (1.4)) is required.) The proof relies on the bounded Hölder theorem established in (3.1) but it is quite complicated and I shall not present it here. It can be greatly simplified if unrestricted solvability is assumed.

It was recognized by Tversky [36] that additive conjoint measurement can be profitably applied to metric representation of similarity orderings. Suppose that we have a dissimilarity ordering  $\succeq$  on  $A \times A$ , as in section 3, but that the set  $A$  is endowed with product structure,  $A = \prod_{i=1}^n A_i$ . If the ordering on  $A \times A$  is to be represented by a Euclidean metric  $d$ , with the sets  $A_i$  as a complete set of orthogonal coordinates, then there are functions  $\phi_i$  on  $A_i$  such that

$$d(a, b) = \left[ \sum_{i=1}^n |\phi_i(a_i) - \phi_i(b_i)|^2 \right]^{1/2}.$$

---

\* A topological version of this theorem was published by Debreu [9]. The present version is due to R. D. Luce, and is taken from material for a book in preparation by R. D. Luce, P. Suppes, A. Tversky, and the present author.

This equation can be generalized in 2 ways:

$$d(a,b) = F \left[ \sum_{i=1}^n \phi_i(a_i, b_i) \right] \quad (1)$$

$$d(a,b) = F \left[ |\phi_1(a_1) - \phi_1(b_1)|, \dots, |\phi_n(a_n) - \phi_n(b_n)| \right]. \quad (2)$$

In (1),  $F$  is a strictly increasing function of 1 variable, while in (2),  $F$  is strictly increasing in each of the  $n$  variables. Equation (1) specifies that dimensions combine additively to determine dissimilarity, while (2) specifies that the contribution of any one dimension can be represented by absolute differences of scale values.

If  $d$  is not required to be a metric, then Equation (1) is simply the equation of additive conjoint measurement, in  $n$  variables. Similarly, when  $d$  need not be a metric, then the absolute difference representation, on any one dimension, can be analyzed by methods very close to those of additive conjoint measurement in 2 variables. (Absolute difference representations have been extensively studied by Pfanzagl [28].)

If we combine the conditions for  $n$ -factor additivity (across dimensions), for 2-factor additivity (within each dimension), and for a metric with additive segments (section 3 above), then, as was shown by Tversky, the metric  $d$  is constrained to the form,

$$d(a,b) = F^{-1} \left[ \sum_{i=1}^n F(|\phi_i(a_i) - \phi_i(b_i)|) \right] \quad (3)$$

where  $F$  is an increasing function, satisfying  $F(\alpha + \beta) \geq F(\alpha) + F(\beta)$ .

If  $F(\alpha) = \alpha^r$ ,  $1 \leq r < \infty$ , then  $d$  is the Minkowski  $r$ -metric. I do not know what other functions  $F$ , if any, yield a metric with additive segments satisfying Equation (3); but if the additive segments are required to lie on algebraic straight lines (in the coordinates  $\phi_1(a_1), \dots, \phi_n(a_n)$ ), then the  $r^{\text{th}}$  power is the only solution. These considerations lead to an axiomatization of the Minkowski  $r$ -metrics in terms of the primitives  $A = \prod_{i=1}^n A_i$  and  $\geq$  on  $A \times A$ . The Euclidean case,  $r = 2$ , can be distinguished by numerous properties, for example, the fact that all rotations are isometries.

If we let  $F(\alpha) = e^{r\alpha} - 1$ ,  $r > 0$ , then (3) yields a metric, but only points differing in exactly one dimension can be joined by an additive segment. The geometry of this "exponential metric" seems not to have been studied.

#### 4.3 Simple polynomials

We now return to consideration of polynomial combination rules more general than additivity. The independence or sign-dependence properties that are logically necessary for a given polynomial are usually not sufficient, even if appropriate solvability and Archimedean conditions are assumed. In this respect, the pure additive and pure multiplicative rules are exceptional. For one class of polynomial combination rules, called simple polynomials, there is a general schema for finding a sufficient set of axioms. This class is defined as follows: (i) single variables are simple polynomials; (ii) if  $f_1$  and  $f_2$  are simple polynomials with disjoint variables, then  $f_1 + f_2$  and  $f_1 f_2$  are simple poly-

nomials; (iii) no polynomials are simple except by virtue of (i) and (ii). More formally:

DEFINITION 4.1 Let  $F[Y]$  be a ring of polynomials in the indeterminates  $Y$ . Then  $S[Y]$  is the smallest subset of  $F[Y]$  such that

(i)  $Y \subset S[Y]$

(ii) if  $Y_1, Y_2 \subset Y$ , with  $Y_1 \cap Y_2$  empty, then for any  $f_1 \in S[Y_1]$  and  $f_2 \in S[Y_2]$ ,  $f_1 + f_2$  and  $f_1 f_2 \in S[Y]$ .

The elements of  $S[Y]$  are the *simple polynomials* of  $F[Y]$ .

To axiomatize polynomial conjoint measurement, relative to a simple polynomial  $f$ , for  $(A_1, \dots, A_n, \succeq)$ , we proceed to introduce a series of addition and multiplication operations in the set  $A/\sim$ . An addition operation is introduced for each decomposition of a simple component of  $f$  as the sum of smaller simple components with nonoverlapping variables, and similarly for multiplications. The manner of introducing these operations is the same as in section (1.4), e.g.,  $(b_1, a_2^0) + (a_1^0, b_2) = (b_1, b_2)$ , where  $a^0 = (a_1^0, a_2^0)$  is the origin (or unit, for multiplication). However, one must be careful to keep the origin the same for all additions, the unit the same for all multiplications, and to choose the origin as a multiplicative zero. In addition to suitable sign-dependence, solvability and Archimedean conditions, three classes of axioms need to be introduced:

(i) appropriate cancellation conditions, like Axiom 3 of (1.4), that guarantee that the operations introduced are well-defined, commutative, and associative;

(ii) conditions guaranteeing that all the addition operations have the same effect, as do all the multiplications;

(iii) a condition that is used to prove distributivity of multiplication over addition.

Of course, all these axioms may not be logically independent, so that some of them may be eliminable in any given instance.

Rather than give complete abstract details of this general schema, I shall sketch an illustration for 4-factor conjoint measurement, relative to the polynomial  $x_1x_2 + x_3x_4$ .

We choose a suitable origin,  $a^0 = (a_1^0, a_2^0, a_3^0, a_4^0)$  and a suitable unit,  $a^1 = (a_1^1, a_2^1, a_3^0, a_4^0) - (a_1^0, a_2^0, a_3^1, a_4^1)$ . Addition is defined by

$$(b_1, b_2, a_3^0, a_4^0) + (a_1^0, a_2^0, b_3, b_4) = (b_1, b_2, b_3, b_4).$$

Two different multiplications are defined by,

$$(b_1, a_2^1, a_3^0, a_4^0) \cdot (a_1^1, b_2, a_3^0, a_4^0) = (b_1, b_2, a_3^0, a_4^0)$$

$$(a_1^0, a_2^0, b_3, a_4^1) \cdot (a_1^0, a_2^0, a_3^1, b_4) = (a_1^0, a_2^0, b_3, b_4).$$

These definitions are chosen so that any element  $b$  of  $A/\sim$  satisfies

$$(b_1, b_2, b_3, b_4) = (b_1, a_2^1, a_3^0, a_4^0) \cdot (a_1^1, b_2, a_3^0, a_4^0) + (a_1^0, a_2^0, b_3, a_4^1) \cdot (a_1^0, a_2^0, a_3^1, b_4).$$

Once we have constructed a ring isomorphism  $\phi$  of  $(A/\sim, +, \cdot, \geq)$  into

$(Re, +, \cdot, \geq)$ , we can define  $\phi_1(b_1) = \phi(b_1, a_2^1, a_3^0, a_4^0)$ , etc., and obtain,

from the previous equation and the isomorphism property, the desired relation

$$\phi(b) = \phi_1(b_1) + \phi_2(b_2) + \phi_3(b_3) + \phi_4(b_4).$$

Unlike the simple additive case, the origin  $a^0$  and unit  $a^1$  must be chosen with some care, and the solvability and cancellation conditions must be formulated to take note of exceptions. In particular, we start by assuming that  $A_1 \times A_2$  and  $A_3 \times A_4$  are mutually independent, while  $A_1$  and  $A_2$ , as well as  $A_3$  and  $A_4$ , are mutually sign-dependent. For  $i = 1, 2, 3, 4$ ,  $a_i^0$  must be chosen in the 0-class determined by sign-dependence, while  $a_i^1$  must be chosen outside the 0-class. This amounts to choosing the additive origin to be the multiplicative zero, and the multiplication to be nontrivial. The solvability and cancellation conditions must be formulated with due care for division by zero. Lastly, some delicacy is required in designating + and - signs, relative to sign-dependence. Once all this is done, and suitable conditions of type (ii) and (iii) are imposed, guaranteeing that the multiplications coincide and are distributive over addition, we obtain an ordered ring, as required.

To illustrate the derivation of type (ii) and (iii) axioms, I indicate one that guarantees distributivity of multiplication over addition. Represent 3 arbitrary equivalence classes by

$$\begin{aligned} b &= (b_1, a_2^1, a_3^0, a_4^0) \\ c &= (a_1^0, a_2^0, b_3, a_4^1) \\ d &= (a_1^1, b_2, a_3^0, a_4^0) - (a_1^0, a_2^0, a_3^1, b_4). \end{aligned}$$

$$\text{Then } b + d + c + d = (b_1, b_2, b_3, b_4)$$

$$b + c = (b_1, a_2^1, b_3, a_4^1).$$

Choose  $c_1$  such that  $(b_1, a_2^1, b_3, a_4^1) = (c_1, a_2^1, a_3^0, a_4^0)$ . Then

$$(b + c) + d = (c_1, b_2, a_3^0, a_4^0).$$

Thus, the required axiom is one that permits us to infer, from the equivalences

$$(a_1^1, b_2, a_3^0, a_4^0) \sim (a_1^0, a_2^0, a_3^1, b_4)$$

$$(b_1, a_2^1, b_3, a_4^1) \sim (c_1, a_2^1, a_3^0, a_4^0)$$

the conclusion

$$(b_1, b_2, b_3, b_4) \sim (c_1, b_2, a_3^0, a_4^0)$$

i.e.,  $b \cdot d + c \cdot d = (b + c) \cdot d$ . With some extra trouble, the required axiom can be formulated as a general condition, independent of the choice of the  $a_1^0, a_3^1$ , and logically necessary for the desired polynomial combination rule.

Finally, note that all the axioms introduced, except for solvability, turn out to be logically necessary conditions for the given polynomial combination rule. It would be interesting to reformulate the above treatment of polynomial measurement with restricted solvability, obtaining only a subset of an ordered ring, but this remains an open problem, which may involve considerable technical difficulty.

## 5. The measurement of color

### 5.1 Metameric matches and vectorial representation

The psychological laws on which color measurement is based were clearly enunciated by H. Grassman in 1853 [12], and bear his name. These laws, and the measurement techniques based on them, are of particular interest, because of their simplicity and beauty, and because they involve an unusual blend of physics, physiology, and psychology. The empirical basis of color measurement involves a physically defined binary operation, *additive color mixture*, and a psychological equivalence relation, *metamerism*. Grassman's laws, which relate these primitives, have clearcut physiological implications (Brindley [5], pp. 198-218). Thus, color measurement has long been the point of departure for physiological and psychological color theories.

The measurement representation for colors involves vectors over the real numbers, rather than real numbers alone. International standards for the vectorial representation of colors were adopted in 1931 by the International Commission on Illumination (ICI) [15]. Extensive discussions of these color measurement standards may be found in the works of Wright [37] and Stiles [32].

In this first section, we assume that the stimulus whose color is to be measured consists of a small, homogeneous patch of light, viewed under standardized conditions. Such a stimulus is specified by a function giving its energy\*, or energy density, for each wavelength in

---

\* The term "energy" is used in a broad sense; depending on the nature of the stimulus, various measures derived from energy may be more appropriate, e.g., power, power per unit area of source, etc. For any change of units the nonnegative measures that correspond to color stimuli need only be altered by a suitable constant factor.

the visible portion of the electromagnetic spectrum (wavelengths from  $4 \times 10^{-7}$  to  $7 \times 10^{-7}$  meters, approximately). Thus, a color stimulus may be considered to correspond to a nonnegative (energy) measure defined on the Borel subsets of a real interval.

*Additive color mixture* means summation of the energy in the mixed stimuli. For example, if stimuli  $b$  and  $c$  are produced by illuminating the same portion of a screen with light from two different projectors, then the mixture,  $b + c$ , is produced by turning both projectors on at once. In fact, the countably additive real-valued set functions on the Borel subsets of the visible spectrum form a vector space over the reals, which we denote  $B$ ; the nonnegative elements of  $B$ , which correspond exactly to the possible specifications of color stimuli, form a convex cone in  $B$ , denoted  $C$ . Additive color mixture corresponds to vector addition in  $C$ .

Two distinct elements of  $C$  may correspond to color stimuli that look alike in color. We say that such stimuli are a *metameric match*, or more simply, are *metamers*. We denote the relation of metamerism by  $\sim$ . With suitable experimental methods, for a normal observer,  $\sim$  can be considered to be an equivalence relation on  $C$ , to a very high degree of approximation. There are many examples of metameric matches: for instance, a stimulus with a "bimodal" energy distribution, with most of the energy in the "red" and "green" parts of the visible spectrum, is metameric to an appropriate "unimodal" stimulus with most of its energy concentrated in the "yellow" part of the spectrum.

Let  $M$  be the set of differences of metameric pairs, i.e.,

$$M = \{d \mid d \in B, \text{ and for some } b, c \in C, \text{ with } b \sim c, d = b - c\}.$$

The content of *Grassman's third law* is that  $M$  is a linear subspace of  $B$ . This experimental finding has been confirmed in modern studies to a high degree of approximation, over a wide range of conditions (see Brindley, [5], p. 211). We can now state succinctly the classical experimental law of color mixture, the *Law of Trichromacy* (*Grassman's first law*):

For a normal observer,  $\dim B/M = 3$ .

An observer is called dichromatic if  $\dim B/M = 2$ , and monochromatic (totally color blind) if  $\dim B/M = 1$ .

Standardized systems of color measurement employ a convenient basis for  $B/M$ . For  $b \in C$ , the coordinates of  $b + M$  relative to the basis in  $B/M$  are called *tristimulus coordinates* of  $b$ . Thus, two stimuli have the same tristimulus coordinates if and only if they are metameric. The ICI standards specify tristimulus coordinates for an average *standard observer*, for approximately monochromatic stimuli (point measures). The tristimulus coordinates for more general stimuli are computed by approximating these as sums of monochromatic stimuli.

Another useful set of coordinates is obtained by regarding the tristimulus coordinates for  $b$  as homogeneous (projective) coordinates for the one-dimensional subspace generated by  $b + M$ . If these homogeneous coordinates are normalized so they sum to 1, they are called *chromaticity coordinates*. Two stimuli have the same chromaticity coordinates if a scalar multiple (change in overall energy level) of one of them is metameric to the other.

## 5.2 Photopigments

The biological effects of light are mediated through absorption by photopigments. The absorbing properties of a photopigment are specified by a spectral absorptance function  $p$ . For wavelength  $\lambda$ ,  $p(\lambda)$  is the fraction of the incident energy at wavelength  $\lambda$  which is absorbed by the pigments and converted into electrochemical energy. For a photopigment in human visual receptors, it is convenient to include in  $p(\lambda)$  the wavelength-dependent alterations in the stimulus between the point where it is specified by a measure and the point where it is absorbed by a receptor, e.g., reflection at the cornea, absorption and scattering in the ocular media. If this is done, then the average number of quanta absorbed and converted into electrochemical energy, from a stimulus  $b$ , by a photopigment  $p$ , is proportional to

$$\int \lambda p(\lambda) db(\lambda).$$

The integral is taken over the visible spectrum; the factor  $\lambda$  is introduced because average quanta/unit energy is proportional to wavelength. Thus, a photopigment  $p$  defines a linear functional on  $C$ , and by the natural extension, on  $B$  as well. (Departures from linearity occur insofar as the photopigment is appreciably depleted by the ensuing photochemical reaction).

Grassman's laws would be explained if we assume that there are 3 linearly independent photopigments,  $p_1, p_2, p_3$ , which mediate color vision. The intersection of the null-spaces of the  $p_i$  would be exactly  $M$ . The  $p_i$  define linear functionals  $\bar{p}_i$  on  $B/M$ , which span the dual space of  $B/M$ . Their dual basis yields a preferred coordinate system in  $B/M$ .

Relative to this basis, the  $i^{\text{th}}$  tristimulus coordinate of  $b$  is precisely  $p_i(b) = \int \lambda p_i(\lambda) db(\lambda)$ , the effective quantal absorption of stimulus  $b$  by  $p_i$ . One may hope to account for various other properties of color (for example, color discriminability or perceived hue) in a simple way in terms of those coordinates, which represent the basic physiological responses (the "Grundempfindungen" of Helmholtz). This is the thesis of the Young-Helmholtz theory of color vision.

An important variant of the 3-pigment hypothesis postulates  $n$  linearly independent photopigments,  $p_1, \dots, p_n$ ,  $n \geq 3$ , which in turn contribute linearly to 3 independent outputs  $q_1, q_2, q_3$ , where

$$q_i = \sum_{j=1}^n a_{ij} p_j, \quad i = 1, 2, 3.$$

The  $q_i$  are again linear functionals on  $B$ , whose null-spaces intersect in  $M$ , and which induce a preferred coordinate system in  $B/M$ . This variant allows a stage of "recoding", represented by the linear transformation  $(a_{ij})$ , between photopigment absorption and the basic outputs that determine other aspects of color vision, such as discriminability, subjective appearance, etc. This sort of recoding is a feature of many color theories, particularly, the Hering opponent-colors theory as quantified by Hurvich and Jameson [14].

There is strong evidence that, if such a recoding does take place, nevertheless, only 3 independent photopigments are involved, i.e.,  $n = 3$ . The key to the argument is the finding that metameric matches are not broken down by moderate adaptation to colored lights. If  $b \sim c$  for normal adaptation, then the appearance of both  $b$  and  $c$  changes after adaptation to colored light, but except for extreme adaptations,  $b$  and  $c$  still match in color.

It is generally assumed that the effect of adaptation involves (among other things) bleaching of the photopigments, i.e., the spectral absorptance function  $p_i$  is multiplied by a constant  $t_i$ , where  $1 - t_i$  represents the fraction bleached. Let  $P$  be the space of linear functionals on  $B$  spanned by  $p_1, \dots, p_n$ . Then the equations

$$Tp_i = t_i p_i \quad i = 1, \dots, n$$

define a linear operator  $T$  of  $P$  onto itself, with eigenvectors  $p_i$ . Let  $Q$  be the subspace of  $P$  spanned by  $q_1, q_2, q_3$ . Dimensionality considerations show that  $Q$  consists of all the linear functionals on  $B$  which vanish on  $M$ . Since adaptation leaves metameric pairs invariant,  $b \sim c$  implies that  $Tq_i(b) = Tq_i(c)$  for  $i = 1, 2, 3$ ; that is, if  $b - c \in M$ , then  $Tq_i(b - c) = 0$ ,  $i = 1, 2, 3$ . Hence,  $Tq_i \in Q$ ,  $i = 1, 2, 3$ , and it follows that the subspace  $Q$  is invariant under all transformations of form  $T$ . This implies that  $Q$  is spanned by 3 of the vectors  $p_1, \dots, p_n$ .<sup>\*</sup> Hence, so far as mediation of color vision is concerned, there are exactly 3 independent photopigments.

Recently, improved spectrophotometric techniques have yielded direct evidence that there are 3 photopigments in human retinal cones (Rushton [29]; Marks, Dobelle, & MacNichol [26]). The question of whether significant recoding of the 3 photopigment outputs takes place is still a key one for color theory, but seems not to be decidable on the basis of color-matching data alone.

---

\* If the  $t_i$  are distinct, then the minimum polynomial of  $T$ , acting in  $Q$ , must have form  $(x - t_{i_1})(x - t_{i_2})(x - t_{i_3})$ . Since the only eigenvectors associated with  $t_{i_k}$  are multiples of  $p_{i_k}$ , it follows that  $p_{i_1}, p_{i_2}, p_{i_3} \in Q$ .

### 5.3 Color appearance

Tristimulus coordinates, determined by fixing an arbitrary basis for B/M, convey no information about the phenomenal appearance of color. Much of the literature on psychology and physiology of color is devoted to establishing a preferred (linear or curvilinear) coordinate system in B/M, in terms of which color appearance, including color discriminability and similarity, can be accounted for, and to characterizing changes of color appearance, correlated with changes in viewing conditions, as transformations in these preferred coordinates. As was shown in the previous section, the photopigments provide one system of preferred coordinates, but the possibility of recoding needs to be considered.

The most popular color theory involving recoding is the opponent-colors theory of Hering. This was quantified by Hurvich and Jameson [14]. It is based on the observation that the qualities of color appearance can be grouped into 3 pairs, red-green, yellow-blue, and white-black. Any color partakes of at most one quality from each pair. The pairs are *opponent*, in the sense that additive mixture produces cancellation: if  $b$  looks red and  $c$  looks green, then, under the same viewing conditions,  $b + c$  looks either less red than  $b$  or less green than  $c$  or neither red nor green. This suggests that the three recoded outputs,  $q_1, q_2, q_3$  of the previous section, consist of a red-green output [ $q_1(b) > 0$  if  $b$  looks red,  $< 0$  if  $b$  looks green], a yellow-blue output, and a white-black output. These outputs depend on the photopigment absorptions,

as indicated in the linear equations of the previous section, but also on other aspects of viewing conditions, particularly, the presence of other stimuli in adjacent parts of the visual field.

Quantitative versions of opponent-colors theory give a good account not only of the basic opponent-color qualities, but of many other aspects of color appearance. For example, Euclidean distances calculated in terms of differences in the red-green and yellow-blue coordinates, give a reasonably good account of color similarity for monochromatic stimuli of constant brightness (Krantz [16]). This example shows how color measurement makes contact with some of the more general psychological measurement ideas presented in earlier lectures, in particular, with problems of measurement of similarity. Color measurement also makes contact with polynomial measurement; for example, Hurvich and Jameson postulated interrelations among some of the subjective qualities of color that follow polynomial combination rules.

The facts of color appearance, especially the existence of opponent pairs of qualities such as red-green, etc., and the usefulness of opponent-pairs as explanations for other color phenomena, make it relatively certain that opponent-color recoding is the proper basis for color theory, although many quantitative details remain to be worked out. Clearcut physiological evidence for opponent-color recoding has been obtained from microelectrode recording of neural activity in the monkey's visual system, by DeValois and his coworkers [10]. The exact physiological mechanisms remain a mystery.

In concluding the study of color theory and color appearance, I should like to touch on the problem of dependence of color appearance on viewing conditions. This dependence is quite dramatic. For instance, a stimulus that looks reddish-yellow under "normal" conditions may look yellow-green after exposing the eye to a red adapting light, or in the presence of a bright red stimulus simultaneously shown in an adjacent part of the visual field. The most precise tool for studying these changes is *cross-context matching*. If we denote different spatio-temporal contexts by  $\sigma, \tau$  etc., and denote by  $b^\sigma$  stimulus  $b$  viewed in context  $\sigma$ , then we can define  $b^\sigma \sim c^\tau$  to mean that stimulus  $b$ , viewed in context  $\sigma$ , has the same color as stimulus  $c$ , viewed in context  $\tau$ . With appropriate experimental methods,  $\sim$  can be regarded as an equivalence relation.  $b^\sigma \sim c^\sigma$  means that  $b$  is metameric to  $c$ , and, as indicated earlier, it has been found that  $b^\sigma \sim c^\sigma$  implies  $b^\tau \sim c^\tau$  for a wide range of contexts  $\tau$ .

A pair of contexts,  $\sigma, \tau$  defines a function:  $f_{\sigma, \tau}(c) = b$  iff  $b^\sigma \sim c^\tau$ . This is defined for every  $c \in C$  whose appearance, in context  $\tau$ , can be matched by the appearance of some stimulus in context  $\sigma$ . Also,  $f_{\sigma, \tau}$  is well-defined on the corresponding subset of  $B/M$ , and its range can be considered a subset of  $B/M$ . The study of context effects on color appearance is thus reduced to study of the properties of vector-valued functions of vectors, of form  $f_{\sigma, \tau}$ .

In the first section above, it was pointed out that one effect of exposure to an adapting light is likely to be bleaching of photopigments,  $p_i \rightarrow t_i p_i$ . If we postulate that equal appearance corresponds to equal

photopigment outputs, then the corresponding functions  $f_{\sigma,\tau}$  representing changes in this effect of adaptation must be all representable by diagonal matrices in the coordinates corresponding to the dual basis of the  $p_i$ . This is called the von Kries coefficient law [27]. During the 1950's, several unsuccessful attempts were made to use empirical determinations of the functions  $f_{\sigma,\tau}$  to determine the required coordinate system, and hence, to find the photopigment absorptance curves ([6],[25]). However, the functions  $f_{\sigma,\tau}$  seem much better represented by a linear transformation plus a translation than by linear transformations alone. A paper by Krantz [19] provides a general theoretical framework for study of functions defined by cross-context matching, and yields the prediction of linear transformation-plus-translation as a special case in B/M.

# References

- [1] Aczél, J., Pickert, G., & Rado, F. Nomogramme, Gewebe, und Quasigruppen. *Mathematica (Cluj)*, 1960, 2(25), 5-24 (fasc. 1).
- [2] Beals, R. & Krantz, D. H. Metrics and geodesics induced by order relations. Submitted for publication, 1967. Mimeographed preprints available from R. Beals, University of Chicago, or D. Krantz, University of Michigan.
- [3] Birkhoff, G. *Lattice Theory*. (3rd. Edition) Amer. Math. Soc. Colloq. Pub. No. XXV, 1967.
- [4] Blumenthal, L. *A Modern View of Geometry*. San Francisco: W. H. Freeman Co., 1961.
- [5] Brindley, G. S. *Physiology of the Retina and Visual Pathway*. London: Edward Arnold Ltd., 1960.
- [6] Burnham, R. W., Evans, R. M., & Newhall, S. M. Predictions of color appearance with different adaptation illuminations. *J. Opt. Soc. Amer.*, 1957, 47, 35-42.
- [7] Busemann, H. *The Geometry of Geodesics*. New York: Academic Press Inc., 1955.
- [8] Cliff, N. Adverbs as multipliers. *Psychol. Rev.*, 1959, 66, 27-44.
- [9] Debreu, G. Topological methods in cardinal utility theory. In K. J. Arrow, S. Karlin, & P. Suppes (Eds.). *Mathematical Methods in the Social Sciences*, 1959. Stanford: Stanford University Press, 1960.
- [10] DeValois, R. L., Abramov, I., & Jacobs, G. H. Analysis of response patterns of LGN cells. *J. Opt. Soc. Amer.*, 1966, 56, 966-977.

- [11] Fuchs, L. *Partially Ordered Algebraic Systems*. Reading, Massachusetts: Addison-Wesley Co., 1963.
- [12] Grassman, H. On the theory of compound colours. *Philos. Magazine* (London) 1854, series 4, 7, 254-264. (Trans. from *Pogg. Ann. Physik*, 1853, 89, 69-84.)
- [13] Hölder, O. Die Axiome der Quantität und die Lehre vom Mass. *Berichte Verhand. Königl. Sächs. Gesell. Wiss.* (Leipzig), *Math.-Phys. Cl.*, 1901, 53, 1-64.
- [14] Hurvich, L. M. & Jameson, D. An opponent-process theory of color vision. *Psychol. Rev.*, 1957, 64, 384-404.
- [15] International Commission on Illumination, Proc. 8<sup>th</sup> Session, Cambridge, 1931.
- [16] Krantz, D. H. The scaling of small and large color differences. Ph.D. dissertation, University of Pennsylvania, 1964.
- [17] Krantz, D. H. Conjoint measurement: the Luce-Tukey axiomatization and some extensions. *J. Math. Psychol.*, 1964, 1, 218-277.
- [18] Krantz, D. H. Extensive measurement in semiorders. *Michigan Mathematical Psychology Program Technical Report, MMPP 66-6*, University of Michigan, 1966.
- [19] Krantz, D. H. A theory of context effects based on cross-context matching. *J. Math. Psychol.*, accepted for publication, 1967.
- [20] Levine, M. V. Transformations which render curves parallel. Mimeographed, University of Pennsylvania, 1966.
- [21] Luce, R. D. Semiorders and a theory of utility discrimination. *Econometrica*, 1956, 24, 178-191.
- [22] Luce, R. D. Two extensions of conjoint measurement. *J. Math. Psychol.*, 1966, 3, 348-370.

- [23] Luce, R. D. & Marley, A. A. J. Extensive measurement when concatenation is restricted and maximal elements may exist. Mimeographed, University of Pennsylvania, 1966.
- [24] Luce, R. D. & Tukey, J. W. Simultaneous conjoint measurement: a new type of fundamental measurement. *J. Math. Psychol.*, 1964, 1, 1-27.
- [25] MacAdam, D. L. Chromatic adaptation. *J. Opt. Soc. Amer.*, 1956, 46, 500-513.
- [26] Marks, W. B., Dobelle, W. H., & MacNichol, E. F., Jr. Visual pigments of single primate cones. *Science*, 1964, 143, 1181-1183.
- [27] Nagel, W. *Handbuch der Physiologie des Menschen*, v. 3, p. 205-221. Braunschweig: F. Viewig & Sohn, 1905.
- [28] Pfanzagl, J. *Die axiomatischen Grundlagen einer allgemeinen Theorie des Messens*. Schriftenreihen Statist. Institut Univ. Wien, Neue Folge, Nr. 1. Würzburg: Physica Verlag, 1959.
- [29] Rushton, W. A. H. A cone pigment in the protanope. *J. Physiol.*, 1963, 168, 345-359. Also Baker, H. D. & Rushton, W. A. H. The red-sensitive pigment in normal cones. *J. Physiol.*, 1965, 176, 56-72.
- [30] Scott, D. & Suppes, P. Foundational aspects of theories of measurement. *J. Symbolic Logic*, 1958, 23, 113-128.
- [31] Shepard, R. N. The analysis of proximities: multidimensional scaling with an unknown distance function. I. *Psychometrika*, 1962, 27, 125-140.

- [32] Stiles, W. S. The basic data of colour matching. *Phys. Soc. Yearbook*, Phys. Soc. (London), 1955, 44-65. Reprinted in R. D. Luce, R. R. Bush, & E. Galanter (Eds.). *Readings in Mathematical Psychology*, v. 2. New York: Wiley, 1964.
- [33] Suppes, P. A set of independent axioms for extensive quantities. *Portugaliae Mathematica*, 1951, 10, 163-172.
- [34] Tarski, A. Contributions to the theory of models, I, II, III. *Indagationes mathematicae* 1954, 16, 572-581 and 582-588 and 1955, 17, 56-64.
- [35] Tversky, A. A general theory of polynomial conjoint measurement. *J. Math. Psychol.*, 1967, 4, 1-20.
- [36] Tversky, A. The dimensional representation and the metric structure of similarity data. Mimeographed, Center for Cognitive Studies, Harvard University, 1965.
- [37] Wright, W. D. *The Measurement of Colour*. (3rd Edition) London: Hilger & Watts, 1964.

**COMPUTER SCIENCE**

**by**

**ABRAHAM TAUB**

**at the**

**American Mathematical Society Summer Seminar**

**on the**

**Mathematics of the Decision Sciences**

**Stanford University**

**July - August 1967**

## COMPUTER SCIENCE\*

by A. H. Taub

\* A lecture to be presented at the A.M.S. Summer Seminar on the Mathematics of the Decision Sciences on August 4, 1967.

The modern, stored program digital computer is about twenty years old. Its advent and the remarkable improvements in its components and circuitry have created a deep interest in the theory of Automata and the methodology of problem formulation and solution by such devices. Out of the study of computers and the study of new formulations of various problems there has emerged the field of Computer Science, which includes (but is not limited to) such sub-fields as computer circuitry, machine organization and logical design, numerical analysis, theory of programming, theory of automata and switching theory.

Computer Science is closely related to mathematics; indeed, numerical analysis is a branch of mathematics while many of the problems arising from the design and use of computers are intimately associated with questions in combinatorial mathematics, abstract algebra and symbolic logic. But an even more fundamental relationship also exists. The inherent structure of a computer forces one who successfully studies or uses it to strive for the type of generality, abstraction and close attention to logical detail that is characteristic of mathematical arguments.

A brief review of the logically distinct units of a computer system is useful in understanding the nature and the growth of various sub-fields of Computer Science. Inasmuch as such a device is a general-purpose one and automatic (i.e. independent of the human operator after the solution of a problem starts) it should contain certain main organs: (1) a central processing unit (an arithmetic unit), (2) a memory, (3) a control unit and (4) input-output devices.

The function of the central processing unit (CPU) is to perform various operations on arrays of bits contained in it. These arrays of bits are usually called words and may represent numbers or non-numerical data drawn from many diverse fields. The crucial point is that with modern electronic techniques, the CPU can perform its operations with fantastic speeds,  $10^6$  to  $10^8$  operations per second for some CPU's.

and some operations. Thus if the computer is to be used in a manner commensurate with its abilities, it must have available to it at speeds comparable to its execution speeds a sequence of operators and operands.

Such sequences are kept in the memory of the computer.

There must be an organ in the computer which will automatically call forth the various operators and operands and execute the former. This device is called the control of the computer. The list of distinct operations which the control can execute is called the code of the computer.

These three units are internal to the computer itself. There must also exist devices, the input-output organ, whereby the human operator and the machine can communicate with each other.

The design and construction of devices with the properties listed above has stimulated great interest in physical phenomena which may be exploited to construct components for computers. Thus many computer laboratories have studied and are continuing to study the possibility of using lasers and masers as computer memories. The possibility of using cryogenic techniques in the construction of computers has also been under intensive examination.

The design and construction of computers has also stimulated a great deal of theoretical work ranging from the "practical" aspects of switching theory to the theory of automata. I would characterize the first limit mentioned above as being concerned with problems such as the following: What is the minimum number of elements with the ability to perform prescribed elementary logical operations (for example to accept as input two logical variables  $x$  and  $y$  and to have as an output the single logical variable " $x$  and  $y$ ") needed to construct a device capable of forming a given Boolean function of  $n$  Boolean variables.

Automata theory may be said to have originated with Turing's work <sup>[1]</sup> on the general definition of what is meant by a computing automaton. Present day problems in this logical - mathematical theory are concerned with abstract models of devices whose behavior at a particular instant of time depend not only on the pre-

sent input to the machine, but generally on the entire past history of the device including past inputs. In the main it is essentially a chapter in formal logic and as such has the property characterized by von Neumann<sup>[2]</sup> of being "cut off from the best cultivated portions of mathematics, and forced onto the most difficult part of the mathematical terrain, into combinatorics".

In the paper from which the above quotation was taken, von Neumann predicted that automata theory would differ from formal logic in two respects: "(1) The actual length of 'chains of reasoning', that is, the chains of operations, will have to be considered and (2) The operations of logic....will have to be treated by procedures which allow exceptions with low but non-zero probabilities". He expected that the need to take account of point (2) would bring the theory of automata closer to analysis than to combinatorics. This has not happened as yet. Current research in Automata theory is concerned with problems arising in connection with the use of computers, in contrast to the design of computers. Thus many workers in the field are studying among other topics, the theory of programming languages, algebraic coding theory and pattern recognition.

When computers are used in problem "solving" they must be furnished with programs: lists of instructions which must be scanned by the control and executed. If this were just a linear scanning of the sequence of instructions which remain unchanged in form, then matters would be simple. Programming a problem for the machine would consist in translating a meaningful text from one language (for example, the language of mathematics in which the planner will have conceived the problem) into another language (that one of the code of the computer).

This is not the case. Because computers execute orders with great speeds, it is necessary to use iterative and inductive algorithms for problem solving. Then the relation between the program to the mathematically conceived procedure of solution is not a statical one, that of translation, but highly dynamical: A program stands not simply for its contents at a given time in reference to a given set of memory locations, but more fully for any succession of passages of the control through it with any succession of modified contents to be found by the control there; all of this being determined by all other orders of the sequence

of instructions (in conjunction with that instruction now being executed by the control).

The theory of programming is in part concerned with techniques for providing a dynamic background to control the automatic evolution of a meaning. von Neumann and Goldstine<sup>[3]</sup> who so characterized programming viewed that subject as a new branch of formal logics and indicated methods for mastering various parts of it. Their work and that of a large body of subsequent workers has been devoted to techniques involved in preparing problems for solution by a computer.

Problem oriented programming languages such as ALGOL, FORTRAN and a host of others have been devised. In addition special languages for dealing with special types of problems such as COBOL, SNOBOL and LISP have also been created.

The existence of these languages and the compilers they require as well as the desire to have efficient use of computers has led to the introduction of operating programs, monitoring systems and executive systems. That is, programs that oversee the use of the computing equipment. This development has led to the creation of a special class of programmers who are called system programmers and who have created a branch of programming called systems programming.

Although the original impetus for developing modern computers arose from the need for better, bigger and faster "arithmetic engines", computers are now being increasingly used for such non-numerical tasks as the simulation of various complex systems, the solution of problems involving only complicated logical operations and even the design of new computers and computer systems. The above list of uses is by no means exhaustive and many of the uses have grown into sophisticated bodies of knowledge with an associated theory that is given a label as a sub-field of Computer Science.

Since other lectures will cover some of these fields in detail I shall confine the remainder of my time to problems arising when one attempts to understand the errors involved when one uses a modern computer to obtain numerical solutions of mathematically formulated underlying physical or engineering problems. J. von Neumann and H.H. Goldstine<sup>[4]</sup> have pointed out the following four sources of

errors in obtaining numerical solutions of an underlying physical or engineering problem:

I ERRORS OF FORMULATION. These arise due to the fact that the mathematical formulation of an underlying physical or engineering problem represents such a problem only with certain idealisations, simplifications and neglects. In other words the problem that is actually formulated to be solved mathematically, the problem that is called the rigorous mathematical problem, is in itself an approximation to some other problem.

There are many examples of these approximations. Thus in the theory of fluid mechanics the notion of a perfect fluid, one without viscosity and without heat conductivity, is an approximate representation of a real compressible fluid. In certain circumstances it represents the behavior of a compressible fluid adequately and its importance arises from this fact. However in other cases it is completely inadequate to deal with the physical problem involved and heat conductivity and viscosity have to be admitted into the theory. Even the theory of a viscous heat conducting compressible fluid, which is much more involved than that of a perfect fluid, is inadequate for some problems. These problems require the replacement of the representation of a fluid as a continuum by the representation of a fluid as a collection of molecules, as is done in the kinetic theory of gases. Thus in this one field we find that successively finer representations of the physical problem have to be made and that the mathematical tools needed for making these refinements are quite different in character.

It is the relationship between the errors of formulation and those listed below that I want to stress today.

II ERRORS IN VALUES OF PARAMETERS. The mathematical formulation of a problem, with the idealisations discussed under I, may involve parameters whose values have to be inserted in a numerical computation. However, these parameters may not be known with sufficient accuracy and thus will introduce errors which may then "infect" the solution.

The discussion of the effects of this source of errors leads to the following mathematical problem: Are the solutions of the rigorous mathematical problem (the approximation to the underlying problem) continuous functions of the parameter? Depending on the answer to this question the importance of the errors of this category can be assessed.

III ERRORS OF TRUNCATION. The mathematical problem described under I may involve transcendental functions and operations which will have to be evaluated by use of a finite sequence of elementary arithmetic operations. For example, the exponential function and the trigonometric functions are usually evaluated by the use of polynomial or rational function approximations and integrals and derivatives are evaluated by quadrature formulas and finite difference approximations respectively.

Thus the rigorous mathematical problem which is an approximation to an underlying problem is further approximated for computational reasons. The interaction between these two approximations is not sufficiently stressed by the people who prepare problems for and make use of computers to obtain numerical solutions of problems.

The approximation of one problem by another raises two mathematical questions of classical numerical analysis: (1) the question of convergence and (2) the question of stability. These questions can be illustrated by considering a problem involving ordinary or partial differential equations.

When the differential equation is replaced by a finite difference equation a mesh is introduced over the independent variables. That is, only discrete sets of values of the independent variables are used in the problem. The question arises as to the behavior of the solution of the difference equations as the number of elements in this set increases. One must search for those discrete formulations of the problem which have the property that their solution converges to the solution of the transcendental problem as the number of elements in the discrete set becomes infinite.

The stability problem involves the question as to whether the solution of the differential equations and the approximating difference equations are continuous functions of the initial and boundary conditions. This question is of importance in view of the fact that because of round-off errors, which will be discussed below, it is impossible to satisfy these conditions exactly. Thus one needs some assurance that the errors introduced from this source will not amplify but will indeed decrease in importance, that is, will damp out.

IV ROUND-OFF ERRORS. These errors arise because the "elementary arithmetic operations" of a computing machine are not rigorously and faultlessly performed. Thus in a digital machine real numbers  $x$  and  $y$  are replaced by digital approximations  $\bar{x}$  and  $\bar{y}$  respectively. The quantities  $x \pm y$  may then have corresponding digital representations, that is, we may have

$$\overline{(x \pm y)} = \bar{x} \pm \bar{y}.$$

However as far as products and quotients are concerned, these are governed by the relations

$$\begin{aligned}\overline{(xy)} &= \bar{x} \bar{y} + \eta'(p) = xy + \eta(p) \\ \overline{(x/y)} &= \bar{x}/\bar{y} + \eta'(q) = x/y + \eta(q)\end{aligned}$$

where  $\eta'(p)$  (and  $\eta(p)$ ) and  $\eta'(q)$  (and  $\eta(q)$ ) are the round-off errors of multiplication and division respectively.

Round-off errors are as old as the art of computation itself. If multiplications and divisions were not rounded-off even elementary computations would soon lead to the use of reams of paper. However, with the advent of high-speed computers the great numbers of such arithmetic operations become possible and one must consider the effect these errors have on the validity of the results obtained.

There is as yet no general theorem governing the effect of round-off errors and every problem formulated for machine solution has to be analyzed separately to obtain an estimate of what the round-off error is and what importance it has for the problem at hand.

The problems with which the theory of round-off and truncation errors should be concerned are closely related to the problem posed by Kronecker when he insisted that mathematics be formulated in a manner which involved finite constructions. When we formulate problems for solution by computing machines we are attempting to follow Kronecker's dictum with at least one important modification: namely, arithmetic is not being carried out faultlessly. The saving grace, if any, being that the errors committed in the arithmetic processes are known.

The formulation of such a theory presents a challenge to pure and applied mathematicians as deep and as important as any problems presently being dealt with. Unfortunately many people are unaware of this challenge. Many mathematicians seem to feel that problems connected with computation are concerned solely with arithmetic, a subject whose avoidance was partly responsible for their becoming interested in mathematics, and therefore such problems are to be shunned.

I now wish to turn to the discussion of the interaction between the errors of formulation and the errors of truncation. My thesis is that such an interaction exists and that it may be profitably exploited to obtain easier and more correct computing algorithms for dealing with various problems. I shall illustrate this thesis by examples. Before doing this I should point out that the existence of present day (and future) computing machines make (and will make) the exploitation of this interaction possible in some cases where without such machines it is not feasible to do anything very different from what has been done in the past.

I shall first discuss what may be considered a ridiculous example but what I hope will not long be a ridiculous one. This example is concerned with a method for dealing with problems in the theory of fluid dynamics. That theory deals with

a continuum over which there are various fields defined: the velocity field, the pressure field and the density field. The physicist considers the fluid as a collection of molecules moving about in space in accordance with certain laws of motion and introduces these various fields in terms of averages of other quantities defined for the molecules themselves.

When a problem in fluid dynamics is formulated in terms of a continuum and fields defined over this continuum which satisfy certain differential equations and when these equations are replaced by finite difference equations we are in effect replacing the fluid by a discrete collection of particles. One should then ask the question as to what relation this collection of particles has with the physicist's molecules. Indeed one should ask why go through this chain of approximations at all? Can we not deal with the physicist's molecules directly and won't this give more significant results than the former procedure?

The answer to the second question seems to be that with existing computers we cannot keep track of enough molecules for a long enough time to make it into a feasible method for handling fluid dynamics problems. However this answer needs to be looked at closely. One reason for this is that no sharp estimates are at hand as to when the law of large numbers becomes operative; will an assembly of thousands of molecules behave essentially the same as  $10^{23}$  molecules as far as the central limit theorem of probability is concerned or are  $10^{23}$  molecules really needed?

The work of Nordsieck and Hicks<sup>[5]</sup> on the application of Monte Carlo techniques to the solution of the Boltzmann equation shows that computers may be effectively used in obtaining the molecular velocity distribution under conditions far from equilibrium. Hence computers enable one to deal with many fluid dynamical problems in a quite novel manner.

There are many instances in which discrete problems are approximated by differential equations which in turn are approximated by difference equations and

then solved by use of computing instruments. Thus errors of formulation and truncation are needlessly introduced, for with the advent of modern high-speed computers the original discrete problem could be handled directly.

Many boundary valued differential equation problems originate from variational principles. This fact can be used to obtain simpler and easily solved discrete approximations to the defining equations for the unknown functions. This is another illustration of the fact that the interaction between the formulation process and the truncation process may be exploited to obtain more meaningful and more accurate approximations to problems posed for numerical solution by computers. I shall not discuss this point further but refer you to a paper<sup>[6]</sup> where this approach has been used on Sturm-Liouville differential equations.

I shall devote the remainder of my remarks to the interaction of round-off errors and truncation errors. Consider the problem of solving the equation

$$x = G(x) \quad (1)$$

on a computer. We shall assume that  $x$  is a real scalar variable and  $G(x)$  is a function such that there exists real numbers  $r$  and  $b$  for which

$$|G(x) - r| \leq b |x - r|$$

where

$$0 \leq b < 1$$

This condition insures the convergence to  $r$  of the sequence of  $x_n$ 's defined by

$$x_{n+1} = G(x_n) \quad n = 0, 1, \dots \quad (2)$$

When one wishes to find the numbers  $x$  satisfying equation (1) one may attempt

to do this by generating on a computer the sequence of  $x_n$ 's satisfying equation (2). However because of truncation errors one will attempt to solve an equation of the form

$$V_{n+1} = H(V_n) \quad n = 0, 1, \dots \quad (3)$$

where

$$H(x) = G(x) + \xi(x) \quad (4)$$

and

$$|\xi(x)| \leq a \quad (5)$$

$a$  being a constant. The function  $H(x)$  is some algebraic approximation to the function  $G(x)$  which in turn may be a transcendental function. The function is the truncation error and  $a$  is a bound for it.

Descloix [7] has shown that for any  $V_0$  the sequence  $V_n$  defined by equation (3) is bounded and all its points of accumulation  $V$  satisfy the inequality

$$|V - r| \leq \frac{|a|}{1-b}.$$

He has further shown that the scheme given by equation (3) is the best possible one for solving equation (2) in the following sense: for given  $a$  and  $b$  there exists a function  $H(x)$  for which it is impossible to find an algorithm using only  $H$ ,  $a$  and  $b$  providing closer points of accumulation to  $r$  than the algorithm (3).

Because of round-off errors a computer will not generate the sequence  $V_n$ . If the computer uses fixed point arithmetic it will generate a sequence of integers

$$y_{n+1} = [G(y_n) + \xi_n]_R$$

where  $[ ]_R$  is called a rounding procedure and  $[x]_R$  is any integer-valued function of  $x$  satisfying the inequality

$$|[x]_R - x| < 1.$$

The normal rounding procedure is defined as

$$[x]_N = [x + 0.5].$$

This is the procedure that is usually used in hand computations and is incorporated in many computers. Descloux showed that for any  $y_0$  the sequence of integers defined by

$$y_{n+1} = [G(y_n) + \xi(y_n)]_N$$

there exists an  $N$  such that

$$|y_n - r| \leq \frac{|a|}{1-b} + \frac{1}{2(1-b)} \quad (6)$$

for

$$n > N;$$

furthermore for given  $a$  and  $b$ , there exists a function  $G$  and errors  $\xi$  for which the bound is attained.

Equation (6) then tells us what accuracy can be expected in solving equation (1) by the iterative scheme (2) when given truncation and round-off errors are introduced. He has obtained the corresponding result when floating point arithmetic is used. I shall not give it here. The point I wish to stress is that convergence may be in error by a significant amount.

Let us consider an example. Suppose we wish to solve the equation

$$x = \frac{7}{8} x.$$

The problem is a trivial one but let us suppose that we do not know that the solution is  $x = 0$  but attempt to solve the equation by generating the sequence

$$V_{n+1} = \frac{7}{8} V_n.$$

If we use fixed point arithmetic with the decimal or binary point at the right of the registers we shall generate the sequence of integers

$$y_{n+1} = \left[ \frac{7}{8} y_n + \frac{1}{2} \right]_N.$$

Thus starting with  $y_0 = 8$ , we obtain  $y_1 = 7$ ,  $y_2 = 6$ ,  $y_3 = 5$ ,  $y_4 = 4$ ,  $y_5 = 4$ ,  $y_n = 4$   $n \geq 5$ . Thus the binary machine solution will be  $4 \times 2^{-p}$  where  $p$  is the number of binary bits in the machine representation of numbers. Of course if  $p$  is large this error is tolerable. On the other hand the result

$$x = 4 \times 2^{-p}$$

is due to the fact that

$$4 = \frac{1}{2(1-b)} = \frac{1}{2(1-\frac{7}{8})}.$$

Hence if  $b$  is close to one, the error may very well be intolerable.

Equation (6) furnishes one with a basis for evaluating different truncating schemes. Thus if  $|a| \ll \frac{1}{2}$  for each of two truncating schemes, the error in the result will be mainly determined by the round-off and hence there is no reason for using the more elaborate truncation procedure.

14.

The question as to whether round-off procedures can be devised which lead to better results has also been considered by Descloux. He extended some results of A. Nordsieck and showed that there exists a round-off procedure such that the  $y_n$  satisfying

$$y_{n+1} = y_n + [G(y_n) + \epsilon_n - y_n]_A$$

are such that for any  $y_0$ , there exists an  $N$  such that

$$|y_{n+1} - r| < \frac{a}{1-b} + 1$$

for

$$n > N.$$

That is the round-off error can be reduced to one bit in the last place. The round-off procedure is called Anomalous rounding for it violates one's intuition. Its rules are

$$\begin{aligned} [x]_A : \text{ for } |x| \leq 1 \quad |[x]_A| &\geq x \\ \text{for } |x| \geq 1 \quad |[x]_A| &\leq x. \end{aligned}$$

Thus when  $x$  is small enough one rounds up, that is, increases the error due to rounding. No one has yet been able to extend these results to the case of vector problems. That is, to the case where  $x$  is a vector and  $G(x)$  is a vector valued function. Thus there remains an open important problem in understanding the interaction between round-off errors and truncation errors.

# REFERENCES

- [1] A. Turing; On Computable Numbers with Applications to the Entscheidungsproblem, Proc. London Math. Soc., 42, (ser. 2), pp. 230 - 265 (1936).
- [2] J. von Neumann; General and Logical Theory of Automata, In Cerebral Mechanisms in Behavior, The Hixon Symposium (September 1948, Pasadena), ed. by L. A. Jeffress (John Wiley, New York), pp. 1 - 31. Reprinted in Collected Works of J. von Neumann, Vol V, Pergamon Press, Oxford, England (1963).
- [3] H. H. Goldstine and J. von Neumann; Planning and Coding Problems for an Electronic Computing Instrument, Collected Works of J. von Neumann, Vol. V, pp. 80 - 151 (1963).
- [4] J. von Neumann and H. H. Goldstine; Numerical Inverting of Matrices of High Order, Bull. Am. Math. Soc., 53, pp. 1021 - 1099 (1947); Collected Works of J. von Neumann, Vol. V, Pergamon Press, Oxford, England, pp. 479 - 572 (1963).
- [5] A. Nordsieck and B. L. Hicks, Monte Carlo Evaluation of the Boltzmann Collision Integral, 5th International Symposium on Rarefied Gas Dynamics, Oxford University, England, July 1966.
- [6] C. C. Farrington, R. T. Gregory and A. H. Taub; On the Numerical Solution of Sturm-Liouville Differential Equations; M.T.A.C., XI, pp. 131 - 150 (1957).
- [7] J. Descloux; Note on the Round-off Errors in Iterative Processes. Math of Comp., XVII, pp. 18 - 27 (1963).

**PERCEPTION PROBLEMS**

by

**ANDRZE J. EHRENFUCHT**

at the

**American Mathematical Society Summer Seminar**

on the

**Mathematics of the Decision Sciences**

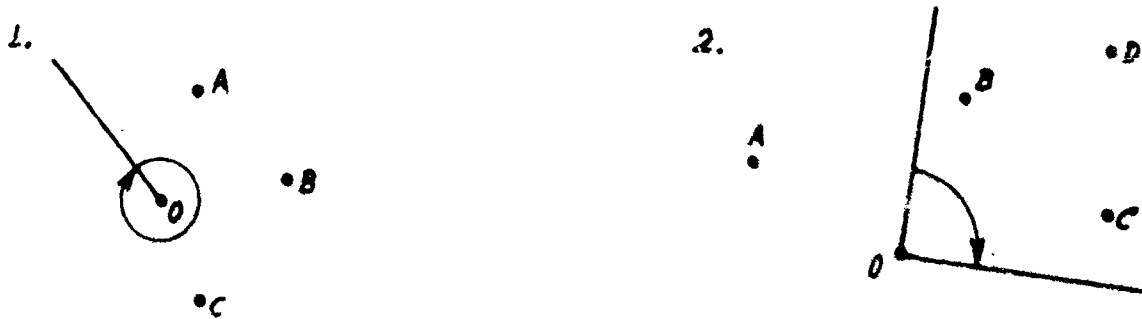
**Stanford University**

**July - August 1967**

TWO DIMENSIONAL VISUAL GEOMETRY

We shall try to describe a geometry based exclusively on the concept "from my position I see an object A to the left of object B". Two dimensional means here, that an observer and all observed objects are on the two dimensional Euclidean plane.

For the objects we shall take marked points on the Euclidean plane (intuitively, they are visible and individually distinguishable objects). The observer will be characterized by his angle of vision which can vary from 0 to 360 degrees. The observer's position on the plane is characterized by a point (the point where he is standing) and the angle equal to his visual angle with the vertex in the point, (the angle covers the area he is observing). The observation from such a position is the permutation (sequence) of marked points which are within his visual angle listed in order from left to right



In the first picture the observer has a visual angle of 360 degrees. From his position marked by O, he made the observation [A, B, C], (he sees A to the left of B, and B to the left of C). In the second

picture the observer's visual angle is 90 degrees and his observation is  $[B, D, C]$ ,  $A$  is not in his field of vision.

Remark. We shall not consider the cases in which the observer's position is in a straight line joining two visible points so that one of them is directly in front of another; but this restriction can be easily avoided.

Let be given a set of observational points  $X$  and a set of marked points  $Y$  from points  $O$  in  $X$  will be called the description of  $Y$ . Such triplet  $(X, Y, \alpha)$  will be called a visual geometry. As the concepts of such a geometry we admit only those which can be defined in terms of the description of  $Y$ . For example, in the geometry where  $X$  is the whole plane and  $\alpha \leq 180$  degrees we can express the fact that point  $a$  lies within the triangle  $B, D, C$  in the following way: in no observation in which  $A, B, C, D$  are visible does  $a$  occupy outer-most left or outer-most right position.

In the case in which  $Y$  has exactly  $n$  elements there are only finitely many possible descriptions of the set  $Y$  (the description is a set of permutations with elements in  $Y$ ). In some cases more exact estimations can be given. For example, when  $X$  is a whole plane and  $\alpha = 360$  degrees the number of different descriptions of  $Y$  has the magnitude of  $n^4$ .

In case when the set  $Y$  consists of all points on the plane (or any dense set of points) then the resulting geometry is affine geometry in the following sense: if the set  $X$  contains at least four non-collinear points, then the only transformation of the plane into itself which preserves the description are affine transformations.

It follows from the last theorem that if we have two configurations of points which are not equivalent in affine geometry we can always add such finite set of points to each configuration so that they will not be equivalent (will have different properties) in respective finite geometries.

This shows that better approximations of affine concepts can be achieved by extending set  $Y$  of visible points and not set  $X$  of observation points, what can be called a type of context sensitivity.

Unclassified

Security Classification

DOCUMENT CONTROL DATA - R&D		
(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)		
1. ORIGINATING ACTIVITY (Corporate author) American Mathematical Society P. O. Box 6248 Providence, Rhode Island 02904		2a. REPORT SECURITY CLASSIFICATION  Unclassified  2b. GROUP
3. REPORT TITLE  Lecture Notes Prepared in Connection with the Summer Seminar on Mathematics of the Decision Sciences		
4. DESCRIPTIVE NOTES (Type of report and inclusive dates)  Lecture notes, July 10-August 11, 1967		
5. AUTHOR(S) (Last name, first name, initial)  Miscellaneous authors; see Table of Contents		
6. REPORT DATE  July 10-August 11, 1967	7a. TOTAL NO. OF PAGES  906	7b. NO. OF PAGES  Unknown
8a. CONTRACT OR GRANT NO.  Nonr(G)-00003-67  a. PROJECT NO.  NR-047-065  c.  d.	9a. ORIGINATOR'S REPORT NUMBER(S)    9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)	
10. AVAILABILITY/LIMITATION NOTICES  U. S. Government agencies may obtain copies of this report directly from DDC. No further copies are available for distribution. Formal publication of these and additional lectures is expected by December 31, 1968		
11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY  Office of Naval Research Logistics and Mathematical Statistics Branch Washington, D. C.	
13. ABSTRACT  The manuscript contains notes of lectures presented at the 1967 Summer Seminar on Mathematics of the Decision Sciences and covers the following subjects: Mathematical Economics, Survey of Mathematical Programming, Reliability Theory, Optimal Stochastic Control, Mathematical Programming, Markovian Decision Processes, Combinatorial Methods, Perception Problems, Networks and Graphs, Integer Programming, Diffusion Approximations in Applied Probability, Branching Processes, Convexity, Measurement and Psychophysics, Computer Science, Control Theory, Learning Theory, Mathematical Linguistics, Mathematical Statistics, Nonlinear Programming, Computational Aspects of Control Theory, Optimal Inventory Control.		

DD FORM 1473  
1 JAN 64

Security Classification

## KEY WORDS

## LINK A

## LINK B

## LINK C

## ROLE

## WT

## ROLE

## WT

## ROLE

## WT

## INSTRUCTIONS

1. **ORIGINATING ACTIVITY:** Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization (*corporate author*) issuing the report.

2a. **REPORT SECURITY CLASSIFICATION:** Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2b. **GROUP:** Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. **REPORT TITLE:** Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. **DESCRIPTIVE NOTES:** If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. **AUTHOR(S):** Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. **REPORT DATE:** Enter the date of the report as day, month, year; or month, year. If more than one date appears on the report, use date of publication.

7a. **TOTAL NUMBER OF PAGES:** The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7b. **NUMBER OF REFERENCES:** Enter the total number of references cited in the report.

8a. **CONTRACT OR GRANT NUMBER:** If appropriate, enter the applicable number of the contract or grant under which the report was written.

8b, 8c, & 8d. **PROJECT NUMBER:** Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9a. **ORIGINATOR'S REPORT NUMBER(S):** Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9b. **OTHER REPORT NUMBER(S):** If the report has been assigned any other report numbers (*either by the originator or by the sponsor*), also enter this number(s).

10. **AVAILABILITY LIMITATION NOTICES:** Enter any limitations on further dissemination of the report, other than those

imposed by security classification, using standard statements such as:

- (1) "Qualified requesters may obtain copies of this report from DDC."
- (2) "Foreign announcement and dissemination of this report by DDC is not authorized."
- (3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through \_\_\_\_\_."
- (4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through \_\_\_\_\_."
- (5) "All distribution of this report is controlled. Qualified DDC users shall request through \_\_\_\_\_."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. **SUPPLEMENTARY NOTES:** Use for additional explanatory notes.

12. **SPONSORING MILITARY ACTIVITY:** Enter the name of the departmental project office or laboratory sponsoring (*paying for*) the research and development. Include address.

13. **ABSTRACT:** Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as (TS), (S), (C), or (U).

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. **KEY WORDS:** Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, rules, and weights is optional.